Improving Large Pose Face Alignment by Regressing 2D and 3D Landmarks Simultaneously and Visibility Refinement

Xu Luo, Pengfei Li, Fuxuan Chen, and Qijun Zhao(✉)

National Key Laboratory of Fundamental Science on Synthetic Vision,
College of Computer Science, Sichuan University, Chengdu, China
qjzhao@scu.edu.cn

Abstract. This paper proposes an improved method for large pose face alignment. Unlike existing methods, the proposed method regresses both 2D and 3D coordinates of facial landmarks simultaneously. It first computes a coarse estimation of the landmarks via a shape regression network (SRN) whose input is only the input image. It then refines the landmarks with another SRN whose input consists of three components: the transformed image, the visible landmark heatmap and the feature map from the first SRN. These components are constructed by a transformation module based on the current estimates of 3D and 2D landmarks. By effectively exploring the 3D property of faces for constraining 2D landmarks and refining their visibility, the proposed method can better align faces under large poses. Extensive experiments on three public databases demonstrate the superiority of the proposed method in large pose face alignment.

Keywords: Face alignment · 3D/2D facial landmarks
Cascaded shape regression · Visible landmark heatmap

1 Introduction

Face alignment, also known as facial landmarks detection, aims at detecting facial key points (such as eye-corners, nose tip, and mouth corners) on face images, which is fundamental to many face-related tasks, e.g., expression recognition, 3D face reconstruction and face recognition. The last decade has witnessed significant progresses in face alignment. With the introduction of cascaded regression [1], many state-of-the-art face alignment methods achieve high precision in detecting the landmarks in frontal and near-frontal (i.e., yaw rotation angles are within ±60°) face images. However, they may still fail in challenging large pose face alignment, due to self-occlusion and unreliable features around invisible landmarks on the face images.

Many recent methods [2–6] use convolutional neural networks (CNN) to learn more effective features rather than using hand-crafted features for detecting

J. Zhou et al. (Eds.): CCBR 2018, LNCS 10996, pp. 349–357, 2018.
https://doi.org/10.1007/978-3-319-97909-0_38

facial landmarks. Some other recent methods resort to 3D face models [7–9] to improve the robustness of facial landmarks detection to large pose variations, from which 2D-based methods suffer. Such 3D-based methods generally fit a 3D morphable face model (3DMM) [11] to the input 2D face image and infer landmarks from the reconstructed 3D face via 3D-to-2D projection. Despite the significant progresses made by CNN-based methods [2–6] and 3D-based methods [7–10], large pose face alignment is still a challenging problem.

In this paper, we propose an improved method to solve the large pose facial landmarks detection problem. Instead of fitting a 3DMM, we directly regress 3D landmarks based on CNN to refine 2D landmarks. It imposes a strong shape constraint to the 2D landmarks. To exclude unreliable features around invisible landmarks, we estimate the visibility of the landmarks based on the obtained 3D coordinates, and generate a visible landmark heatmap that can facilitate the extraction of pose-robust features. Evaluation results on three public benchmark databases with comparison to state-of-the-art methods prove the effectiveness of our proposed method.

2 Related Work

Many methods utilize 3D face alignment to refine 2D face alignment for large pose faces considering the limitation of 2D-based methods in dealing with self-occlusion. Zhu et al. [9] proposed a method called 3D Dense Face Alignment (3DDFA), which generated PNCC map from the obtained 3D face shape and stacked it with the input image as the input to the next stage. Although having well advanced the state-of-the-art of face alignment, like most existing 3D-based methods [7,8], it still has difficulties in dealing with near profile faces because it does not explicitly consider invisible landmarks. Chen et al. [10] refined 2D face landmarks by using 3D landmarks that were regressed from hand-crafted features. These 3D-based methods, regressing either 3DMM parameters or 3D coordinates, compute 2D landmarks via projecting the obtained 3D landmarks onto 2D images. In this paper, instead, we regress directly both 3D and 2D coordinates of the landmarks with learned features, and use the 3D landmarks as a strong shape constraint to refine the 2D landmarks.

Various types of feature maps have been used to assure focusing on the region of interest and extracting more robust features. DAN [5] aims at detecting visible facial contour points and utilizes landmark heatmaps to constrain the region of interest from which features are extracted. However, the heatmaps in DAN do not consider the visibility of landmarks, and would thus lead to unreliable features around invisible landmarks. The PNCC feature maps [9] are obtained by projecting 3D face shapes onto 2D plane via z-buffering. The Z-Buffer representation is, however, not differentiable, preventing end-to-end training. In our work, we utilize the regressed 3D landmarks to estimate the visibility of each landmark and generate heatmaps based on the visible landmarks. This way, we can better ensure that more robust features are learned.

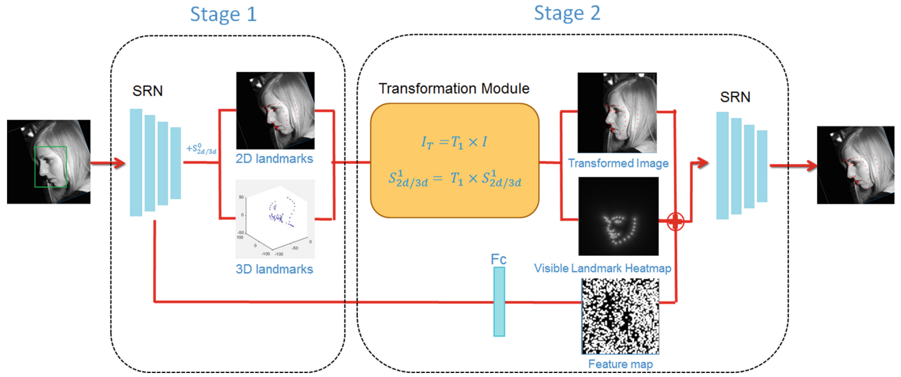

Fig. 1. The main steps in our proposed method, which regresses the landmarks in two stages. The first stage coarsely estimates the residual of 2D and 3D landmarks with respect to the landmarks' initial estimates (i.e., the mean locations in frontal view). The second stage refines the estimated 2D and 3D landmarks by taking the transformed image (I_T), the visible landmark heatmap and the feature map from the prior stage as input, which are generated by the transformation module. In practice, the second stage can be repeated, resulting in a deeper cascade structure, though we implement only two stages in this paper.

3 Proposed Method

3.1 Overview

Figure 1 shows the main steps in our proposed method, which consists of two stages. In the first stage, a Shape Regression Network (SRN) is employed to generate coarse estimates of both 2D and 3D landmarks for the input face image. Unlike general 3D-based face alignment methods that need to fit a 3DMM, we directly regress 3D landmarks and their corresponding semantically consistent 2D landmarks.

In the second stage, another SRN is deployed to refine the estimated 3D and 2D landmarks. To fully explore the knowledge obtained in the first stage, we combine the information from three different sources to form the input of the SRN, specifically, the transformed input image, the heatmap of currently estimated visible 2D landmarks, and a feature map from the first stage SRN. In the transformation module, the transformation applied to the input image as well as the 2D landmarks, and the visibility of 2D landmarks are computed.

3.2 Shape Regression Network

As shown in Fig. 2, the structure of SRN is inspired by the VGG network [12]. While the SRNs in the two stages share similar structure, they differ in their inputs: The input of the first SRN is the original input image; but the input of the second SRN is a combination of the transformed input image, visible landmark

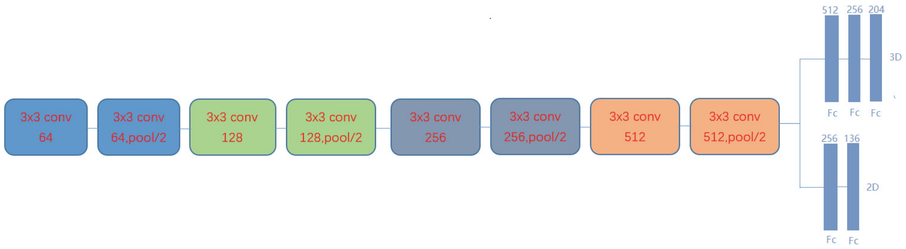

Fig. 2. The inputs of the two SRN in our proposed method are 112 × 112 × 1 and 112 × 112 × 3, respectively.

heatmap and a feature map from the first SRN. These three components are generated in the transformation module and stacked across channel. Each SRN regresses simultaneously 2D and 3D shape $\varDelta S_{2d}$ and $\varDelta S_{3d}$, which are used to update the current estimates of 2D landmarks S_{2d} and 3D landmarks S_{3d}.

3.3 Transformation Module

The transformation module generates the input for the second SRN based on the output of the first SRN. Specifically, it transforms the input image as well as its 2D landmarks to a canonical frontal view via an affine transformation. The parameters involved in the transformation (denoted by T) are estimated by minimizing the error between the transformed 2D landmarks and the mean 2D landmarks on frontal face images ($\hat{S}_{2d}^{F}$):

$$\arg\min_{T} ||\hat{S}_{2d}^{F} - T \times S_{2d}||_2^2. \tag{1}$$

with the computed affine transformation, the original input face image and its 2D landmarks are transformed accordingly with bilinear interpolation.

Since the transformed image is used as input to the second SRN, its regressed shape residuals should be transformed back to the coordinate system of the original input image. Hence, the refined 2D/3D landmarks in the second stage are computed as follows,

$$S_{2d/3d}^{2} = T_2^{-1} \times (T_2 \times S_{2d/3d}^{1} + \varDelta S_{2d/3d}^{2}) \tag{2}$$

where $\varDelta S_{2d/3d}^{2}$ is the output of the SRN of stage 2, T_2^{-1} is the inverse of transform T_2.

Note that the visibility of the landmarks is not considered so far. Fortunately, the estimated 3D landmarks can be used to determine the visibility. Let M denote the weak perspective projection matrix from 3D to 2D. We compute it by minimizing the fitting error between the 3D and 2D landmarks. Given the 3D landmarks and the 3D-to-2D projection matrix, we can compute the visibility of the corresponding 2D landmarks. More detail will be given in the next subsection.

3.4 Visible Landmark Heatmap

To utilize landmark heatmap to improve the quality of extracted features especially for large pose faces, we need to estimate the visibility of each facial landmark based on the corresponding 3D and 2D landmarks. The computation of visibility is proposed in [7]:

$$v = sign(\overrightarrow{N_i} \cdot (\frac{m1}{||m1||} \times \frac{m2}{||m2||})) \tag{3}$$

where $m1$ and $m2$ are, respectively, first row vector and second row vector of the 3D-to-2D projection matrix M, $\overrightarrow{N_i}$ is the normal vector at the landmark i in 3D space, and $sign$ denotes the sign function. Thus, if v is positive, the landmark is visible; otherwise invisible.

After estimating the visibility of each landmark, we utilize the visible landmarks to generate visible landmark heatmap. Landmark heatmap is an image whose pixel intensity has an inverse relationship with the distance between the pixel location and nearest landmark location. The visible landmark heatmap can be computed by

$$H(x, y) = \frac{1}{1 + min_{s_i \in T_k \times S_{2d}^k} ||(x, y) - s_i||_2} \tag{4}$$

where $H(x, y)$ is the intensity of pixel (x, y) in visible landmark heatmap image, $T_k \times S_{2d}^k$ are transformed visible landmarks at regression stage k, s_i is the nearest visible landmark of pixel (x, y).

3.5 Feature Map

The feature map is an image generated by a fully connected layer, whose input is the convolutional feature map of the last pooling layer in SRN. The output size of the fully connected layer is $12,544$, and the output is reshaped to an image $(112 \times 112 \times 1)$. The feature map as a complement to the input facial image and visible landmark heatmap transfers the learned information of prior stage to later stage.

3.6 Loss Function

At each stage, we learn to minimize the 2D and 3D landmarks location error normalized by facial bounding box diagonal lengths. Therefore, our loss function can be written as

$$L = \frac{||T_k^{-1}(T_k S_{2d}^{k-1} + \Delta S_{2d}^{k-1}) - S_{2d}^*||}{d_{2d}} + \frac{||T_k^{-1}(T_k S_{3d}^{k-1} + \Delta S_{3d}^{k-1}) - S_{3d}^*||}{d_{3d}} \tag{5}$$

where S_{2d}^* and S_{3d}^* are ground truth 2D and 3D landmarks, d_{2d} and d_{3d} are the diagonal lengths of the 2D and 3D facial bounding boxes respectively. Note that in the first SRN the input is original face image. Therefore, the loss function of the first stage does not include the transformation T or its inverse T^{-1}.

4 Experiments

4.1 Implementation Details

We train our model with the 300W-LP database [9], which contains 61,225 images of front, middle-front and challenging profile faces together with their 68 ground truth 3D landmarks and their corresponding semantically consistent 2D landmarks. To increase the data diversity, we do data augmentation for the training data by applying mirror, rotation, translation and scaling.

While our model consists of two stages, we first pre-train the first stage, and then train both stages together in an end-to-end manner. We use Adam stochastic optimization [14] to optimize our loss with a learning rate of 0.001 and mini batch size of 64. The method is implemented with Tensorflow 1.4.0. The obtained model can run at 35 fps on a computer with one GeForce GTX 1050Ti.

4.2 Experimental Results

We compare our method with some state-of-the-art methods on three databases: AFLW- 2000-3D [9], Menpo-3D [15], and 300W-Testset-3D [16]. In the experiments, we use the facial bounding boxes generated from ground truth landmarks, and the mean frontal face shape as the initial face shape $S^0_{2d/3d}$.

Table 1. Mean error normalized by bounding box diagonal length on AFLW2000-3D database.

Method	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	Mean
RCPR (300W-LP) [17]	4.26	5.96	13.18	7.80
ESR (300W-LP) [18]	4.60	6.7	12.67	7.99
SDM (300W-LP) [19]	3.67	4.94	9.76	6.12
3DDFA [9]	3.78	4.54	7.93	5.42
3DDFA+SDM [9]	3.43	4.24	7.17	4.94
Chen et al. [10]	3.20	5.48	6.12	4.93
3D-FAN [6]	3.38	4.46	5.59	4.47
Ours (heatmap)	2.97	3.93	5.18	4.02
Ours (visible-heatmap)	**2.97**	**3.85**	**5.09**	**3.97**

Table 2. Mean error normalized by bounding box diagonal length on AFLW2000-3D database (Only visible landmarks are considered).

Method	$[0°, 30°]$	$[30°, 60°]$	$[60°, 90°]$	Mean
DAN [5]	3.07	4.01	8.16	5.08
Ours	**2.92**	**3.36**	**4.12**	**3.46**

AFLW2000-3D is a challenging large pose database containing 2,000 facial images and their annotated ground truth 68 semantic landmarks. We categorize the face images in AFLW2000-3D into three view groups $[0^\circ, 30^\circ], [30^\circ, 60^\circ], [60^\circ, 90^\circ]$ according to their yaw rotation angles. The resulting three groups contain $1,312$, 390 and 298 images, respectively. Table 1 shows the landmark localization errors of the proposed method and the counterpart methods. Obviously, our method achieves the lowest error. In Table 1, we also report the performance of our method when conventional heatmap rather than the visibility-refined heatmap is used. The increased error proves the importance of considering the landmark visibility. Table 2 further compares our method with the latest DAN method. Note that only visible landmarks are considered here for the sake of fair comparison. Again, our method performs better.

Menpo-3D contains 8,955 challenging images with varying illuminations, poses and occlusions. 300W-Testset-3D contains 600 in-the-wild images. We compare our method with Chen et al. [10] and 3D-FAN [6] on these two databases. The results are shown in Table 3, which again demonstrate the superiority of our method in robustly detecting facial landmarks under challenging conditions. Figure 3 shows the landmarks detected by our method on some example images.

Table 3. Mean error normalized by bounding box diagonal length on Menpo-3D and 300W-Testset-3D databases.

Method	300W-Testset-3D	Menpo-3D
Chen et al. [10]	3.38	4.46
3D-FAN [6]	2.83	3.70
Ours	**2.77**	**3.35**

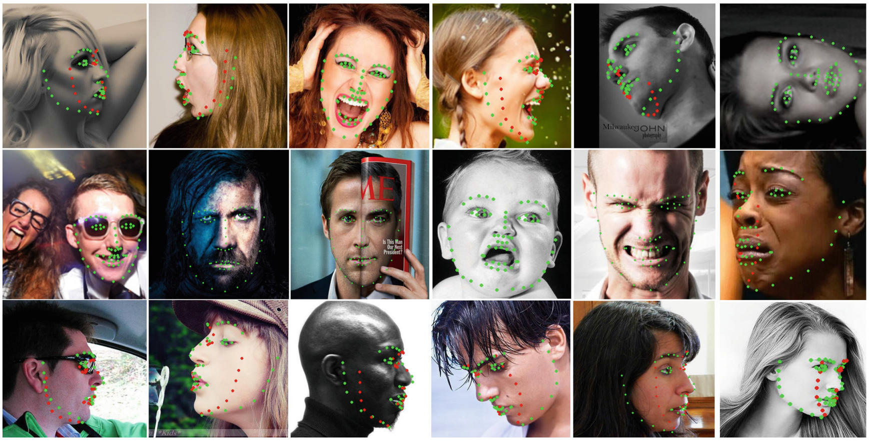

Fig. 3. Landmark detection results of our method on images from AFLW2000-3D (first row), 300W-Testset-3D (second row) and Menpo-3D (third row). Green and red dots show the visible and invisible landmarks, respectively. (Color figure online)

5 Conclusions

In this paper, we propose an improved large pose face alignment method that can locate 2D and 3D facial landmarks simultaneously. Our proposed method effectively explores the 3D property of faces to refine the detected 2D landmarks. Unlike existing methods, our proposed method simultaneously estimates the 2D and 3D coordinates of the facial landmarks, and regularizes the landmark heatmap with the landmark visibility that is determined based on the 3D coordinates. Extensive experiments on challenging databases show that our method is superior to the compared existing methods in challenging large pose face alignment.

Acknowledgements. This work is supported by the National Key Research and Development Program of China (2017YFB0802300) and the National Natural Science Foundation of China (61773270).

References

1. Zhou, S., Comaniciu, D.: Shape regression machine. Inf. Process. Med. Imaging **45**(84), 13–25 (2007)
2. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR, pp. 3476–3483 (2013)
3. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
4. Zhu, S., Li, C., Chen, CL., Tang, X.: Face alignment by coarse-to-fine shape searching. In: CVPR, pp. 4998–5006 (2015)
5. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: a convolutional neural network for robust face alignment. In: CVPRW, pp. 2034–2043 (2017)
6. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D&3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: ICCV, pp. 1021–1030 (2017)
7. Jourabloo, A., Liu, X.: Pose-invariant 3D face alignment. In: ICCV, pp. 3694–3702 (2015)
8. Liu, F., Zeng, D., Zhao, Q., Liu, X.: Joint face alignment and 3D Face reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 545–560. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_33
9. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.: Face alignment across large poses: a 3D solution. In: CVPR, pp. 146–155 (2016)
10. Chen, F., Liu, F., Zhao, Q.: Robust multi-view face alignment based on cascaded 2D/3D face shape regression. In: You, Z., et al. (eds.) CCBR 2016. LNCS, vol. 9967, pp. 40–49. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46654-5_5
11. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint, pp. 1409–1556 (2014)

13. Tuzel, O., Marks, T.K., Tambe, S.: Robust face alignment using a mixture of invariant experts. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 825–841. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_50
14. Kingma, D., Adam, J.: A method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–13 (2014)
15. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: a step closer to the solution. In: CVPRW, pp. 2116–2125 (2017)
16. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: ICCVW, pp. 397–403 (2013)
17. Burgos-Artizzu, X., Perona, P., Dollar, P.: Robust face landmark estimation under occlusion. In: ICCV, pp. 1513–1520 (2013)
18. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014)
19. Xiong, X., Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR, pp. 532–539 (2013)