Modified Stacked Hourglass Networks for Facial Landmarks Detection

Van-Thanh Hoang and Kang-Hyun Jo(✉)

School of Electrical Engineering, University of Ulsan, Ulsan, Korea
thanhhv@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. Facial landmarks detection is a fundamental research topic in computer vision. This topic has been largely improved recently thanks to the development of convolution neural networks (CNN). This paper proposes a modified version of the Stacked Hourglass Network, which is a state-of-the-art architecture for landmark localization. Instead of using the original residual block, this paper uses the λ-residual-block to get more effective features. The proposed network can achieve better result than other state-of-the-art methods on two very challenging 3D facial landmark datasets, Menpo-3D and 300 W.

Keywords: CNN · Hourglass · Facial landmarks · λ-residual block

1 Introduction

Recently, thanks to the advent of Deep Learning and the development of large datasets, many research works have shown results of fantastic accuracy even on the most challenging computer vision tasks. This paper focuses on the task of landmark localization, in particular, on facial landmark localization, also known as face alignment, arguably one of the most heavily researched topics in computer vision over the last decades.

Face alignment or facial landmark estimation is the task of estimating the position of key-points of faces such as eye-corners, mouth corners in an image. As shown in [3], the accurate face alignment can improve the performance of a face verification system, as well as other application such as 3D face modeling, face animation.

Before the advent of deep neural networks, many different techniques have been used for landmark localization. They almost depend on the task in hand. For example, works in human pose estimation was primarily based on sophisticated extensions [23,24,29,31,37] and pictorial structures [9] due to their ability to accommodate a wide spectrum of human poses and also model large appearance changes.

Recently, the Fully Convolutional Neural Network architectures based on the score-map regression have revolutionized the human pose estimation task [14,20,22,25,32,33,35] to produce results of good accuracy even for the very challenging datasets [1]. Thanks to the similarity between estimation human

N. T. Nguyen et al. (Eds.): ACIIDS 2019, LNAI 11432, pp. 648–657, 2019.
https://doi.org/10.1007/978-3-030-14802-7_56

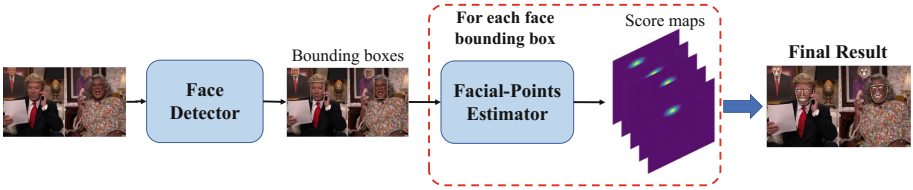


Fig. 1. Overall architecture of the proposed network. From the input image, it uses the faster R-CNN face detector to get the bounding boxes for every face inside the image. Then, for each bounding box, it uses the proposed Facial-Points Estimator to generate the score-maps of all facial points. Finally, it aligns the landmarks of all faces based on the score-maps.

Fig. 2. Example output produced by the proposed Facial Points Estimator. The first left image is the final facial landmarks provided by the max activations across each score map. The other images show sample score maps of some facial key-points (with the original image in behind). From left to right: right jaw, chin, left jaw, left eyebrow, right eye, nose, and mouth.

key-points and facial landmarks tasks, such methods can be freely applied to the problem of facial landmarks alignment.

This paper adopts the top-down approach to align the landmarks for every face in the input image. Firstly, it uses a face detector, which can be MTCNN (Multi-Task Cascade Convolutional Neural Networks) [39] or faster R-CNN (Region-based Convolutional Neural Networks) [16] to get the bounding boxes for every face inside the image. Then, for each bounding box, it crops the input image and uses the proposed Facial-Points Estimator to generate the score-maps of all facial points for the face inside the cropped image. Finally, it aligns the landmarks of all faces based on the score-maps (as can be seen in Fig. 1).

The proposed Facial-Points Estimator is a modified version of the Stacked Hourglass Network [20], which is a state-of-the-art architecture for landmark localization, by replacing the original residual block with the λ-residual block. The examples of output score-maps for some facial key-points are shown in Fig. 2.

2 Related Work

This section reviews related work on face alignment and landmark localization under the two categories: hand-crafted features-based and deep learning-based methods.

Hand-Crafted Features-Based Methods. Zhu et al. [42] proposed the tree structure part model (TSPM) used deformable part-based model for simultaneous detection, pose estimation and landmark localization of face images modeling the face shape in a mixture of trees model. The statistical methods like Active Appearance Models (AAM) [7] and Constrained Local Models (CLM) [8] perform keypoint detection by maximizing the confidence of part locations in a given input image using handcrafted features such as SIFT [21] and HOG. In [2], Asthana et al. proposed a dictionary of the probability response maps followed by linear regression in a CLM framework. Early cascade regression-based methods such as [5,34,41] also used hand-crafted features such as SIFT to capture the appearance of the face image. The major drawback of regression-based methods is their inability to learn models for unconstrained faces in extreme pose.

Deep Learning-Based Methods. Sun et al. [30] used a CNN cascade to regress the facial landmark locations. The work in [39,40] proposed multi-task learning for joint facial landmark localization and attribute classification. The method of [36] and its extended version [34] are within recurrent neural networks. 3DDFA [44] modeled the depth of the face image in a Z-buffer, after which a dense 3D face model was fitted to the image via CNNs. Pose Invariant Face Alignment (PIFA) proposed by Jourabloo et al. [17] predicted the coefficients of 3D to 2D projection matrix via deep cascade regressors.

Landmark Localization. There are many good CNN architectures proposed for key-points estimation task. The stacked hourglass network [20] is the state-of-the-art architecture. Hourglass networks use a stack of 8 very deep hourglass modules to generate the per-pixel labeling task for every key-points. This paper proposes a modified hourglass network to have fewer parameters and achieve better performance.

Face Detection. MTCNN [39] and faster R-CNN [16] are famous in face detection application. Faster R-CNN was proposed by Ren et al. [26] guided by the R-CNN family [10,11] for object detection task. The proposed method uses the Faster R-CNN detector with Resnet-50 backbone for the face detection part.

The proposed method is based on the top-down approach. Firstly, it uses the faster R-CNN face detector to get the bounding boxes for every face inside the image. Then, for each bounding box, Then, for each bounding box, it crops the input image and uses the proposed Facial-Points Estimator to generate the score-maps of all facial points for the face inside the cropped image. Finally, it aligns the landmarks of all faces based on the score-maps.

3 Our Approach

As shown in Fig. 3, the proposed Facial-Points Estimator network has the Residual Networks (ResNet) [13] as the backbone followed by four stacked Modified Hourglass block to generate the score-maps of every facial points.

Deep Residual Network. The ResNet is proposed by He et al. [13] to overcome the degradation problem: "When the network goes deeper and deeper, accuracy

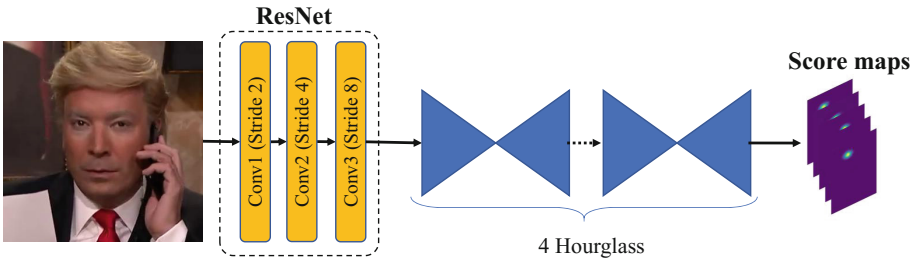


Fig. 3. Architecture of the proposed Facial-Points Estimator. It uses three first blocks of ResNet-50 as backbone. From the input cropped image, it uses ResNet to extract features. Then, these features are fed to four stacked modified hourglasses to generate the score-maps for all facial-points of the face inside the cropped image.

can be saturated and then degrades rapidly". This problem is not caused by overfitting, and adding more layers to a suitably deep model leads to higher error. Their solution is to add the skip connections to make a residual mapping. Figure 4c shows the architecture of an original residual block.

Facial-Points Estimator. The architecture of Facial-Points Estimator is shown in Fig. 3. This proposed network uses three first blocks of ResNet-50 as backbone. From the input cropped image, it extracts features by using ResNet backbone. Then, these features are fed to four stacked modified hourglasses to generate the score-maps for all facial-points of the face inside the cropped image.

Modified Hourglass Block. The architecture of the Modified Hourglass block is shown in Fig. 4a. It uses λ-residual block instead of the original residual block in the main stream. Additionally, it replaces the residual block in the branch stream with a 1×1 Convolution layer to reduce the number of parameters.

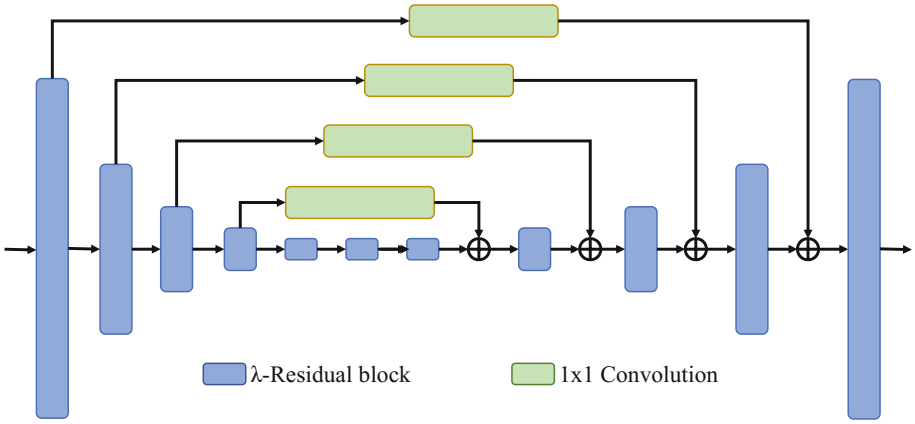
The λ-residual block, as shown in Fig. 4b, used by this network is a little bit different from the original residual block in ResNet. Before the addition, the output of the far previous layer is multiplied by a trainable number λ, it also has another branch with a 1×1 convolution layer to be added before going to the next layer.

4 Experiments

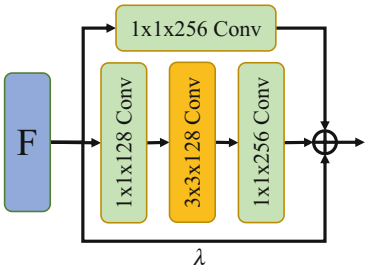
4.1 Implementation Details

The proposed network is trained on the 300W-LP Dataset [43], a synthetically expanded version of 300-W [28]. This dataset provides 3D landmarks allowing for training models and conducting experiments. The 3D annotations are actually the 2D projections of the 3D facial landmarks but for simplicity, we will just call them 3D.

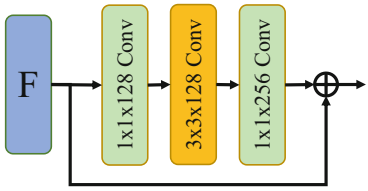
Data augmentation is a simple process to increase number of training images. The augmentation used to train the proposed network are randomly: rotation (from $-50°$ to $+50°$), color jittering, flipping, and scale noise (from 0.5 to 1.2).



(a) Modified Hourglass architecture



(b) λ-Residual block architecture



(c) Original Residual block architecture

Fig. 4. Architecture of Modified Hourglass block and λ-residual block used in the proposed network.

The proposed method is validated on two kinds of dataset. 300-W test set [27] and Menpo-3D [38]. The 300-W dataset consists of the 600 images used for the evaluation purposes of the 300-W Challenge [27]. They are split into two categories: Indoor and Outdoor. The Menpo dataset is a recently introduced for the Menpo Challenger [38] containing 3D landmark annotations for about 9,000 faces from FDDB [15] and ALFW [19].

Similar to [4], this paper use AUC (Area-Under-the-Curve) score as the metric. This score is calculated based on the threshhold of 7% of Normalized Mean Error (NME). The NME is defined as:

$$\text{MNE} = \frac{1}{N}\sum_{k=1}^{N}\frac{\|x_k - y_y\|_2}{d}, \tag{1}$$

Table 1. AUC scores (at NME 7%) of proposed methods and other state-of-the-art methods on the 300W and Menpo datasets. Scores of compared methods are obtained from [4].

Dataset	**Ours**	[4]	[34]	[40]	[41]
Menpo	**68.2%**	67.5%	67.1%	47.9%	60.5%
300W	**74.6%**	66.9%	58.1%	41.7%	55.9%

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 $d = \sqrt{w_{bbox} * h_{bbox}}$.

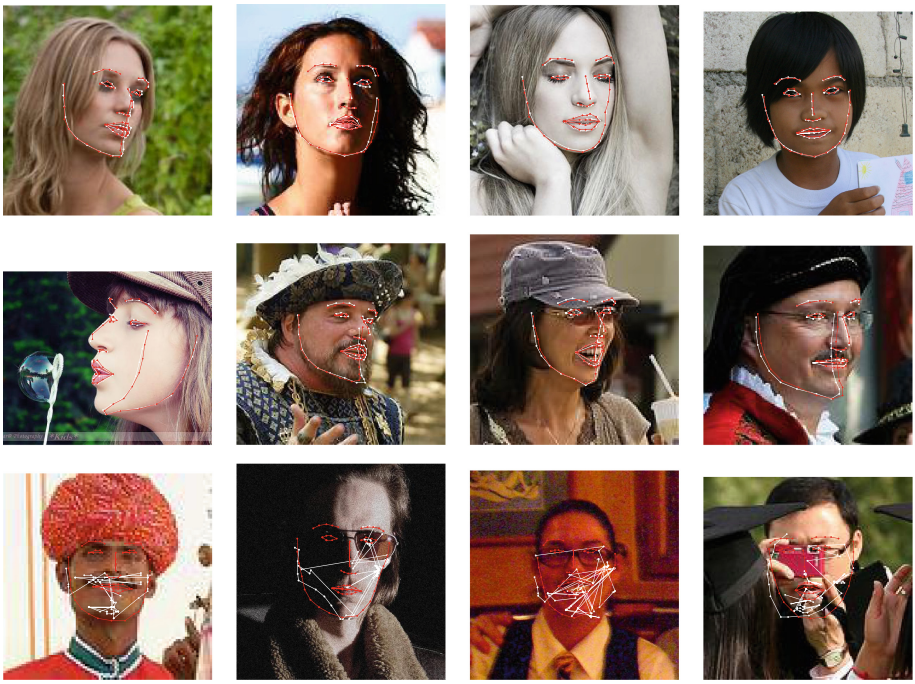


Fig. 5. Qualitative results of the proposed network on the Menpo-3D dataset. Red: ground truth. White: proposed network's predictions. The top two rows are the examples of good results when the predicted facial key-points is similar to the ground truth. The last row is the example of highest-error results when predicted facial landmarks is mixed up. The main reasons are very low resolution, bad lighting, and/or face is behind some objects. (Color figure online)

The proposed network is implemented on MXNet open source deep learning framework [6]. The Adam [18] optimizer implemented by MXNet is used for

training the proposed network. It is trained for 50 epochs on a computer with AMD Ryzen 7 3.60 GHz CPU, NVIDIA 1080Ti GPU device, and 32-GB RAM. The initialized learning rate is set to 4e−5 and then reduced 10 times at 20 and 35 epochs, respectively. The Xavier's initializer [12] is used to initialize the parameters of weighted layers (e.g. Convolution layers). The other settings are: batch size of 24, weight decay of 0.0001, and momentum of 0.9.

4.2 Experiment Results

Table 1 shows the AUC score at NME 7% of the proposed network and other state-of-the-art methods. The proposed network consistently able to outperform the state-of-the-art methods in both two validation datasets. Specially, it has much higher score than compared methods in 300 W dataset.

Some selected visualization results are shown in Fig. 5. As can be seen, most of predicted facial landmarks are similar to the ground truth. But in case of very low resolution, bad lighting, and/or face is behind some objects, the predicted facial landmarks are mixed up. Additionally, these hard cases do not appear in the training dataset much.

5 Conclusion

This paper proposed a facial-points estimator to align the facial landmarks for faces in an image. The proposed network uses the ResNet-50 as backbone, followed by a modified Hourglass networks. Instead of using the original residual block, it uses the λ-residual block to extract better features.

In the future, because the faster R-CNN for face detector and proposed network use the same ResNet as backbone, it is necessary to combine them into one system to have a smaller system, which can run in real-time.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693 (2014)
2. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3444–3451 (2013)
3. Bansal, A., Castillo, C.D., Ranjan, R., Chellappa, R.: The Do's and Don'ts for CNN-based face verification. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2545–2554 (2017)
4. 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: Proceedings of the IEEE International Conference on Computer Vision, p. 4 (2017)
5. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vision **107**(2), 177–190 (2014)

6. Chen, T., et al.: MXNET: a flexible and efficient machine learning library for heterogeneous distributed systems. In: Neural Information Processing Systems, Workshop on Machine Learning Systems (2015)
7. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 681–685 (2001)
8. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recogn. **41**(10), 3054–3067 (2008)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vis. **61**(1), 55–79 (2005)
10. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
12. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 34–50. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_3
15. Jain, V., Learned-Miller, E.: FDDB: a benchmark for face detection in unconstrained settings. Technical report UM-CS-2010-009, University of Massachusetts, Amherst (2010)
16. Jiang, H., Learned-Miller, E.: Face detection with the faster R-CNN. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 650–657. IEEE (2017)
17. Jourabloo, A., Liu, X.: Pose-invariant 3D face alignment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3694–3702 (2015)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
19. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2144–2151. IEEE (2011)
20. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
21. Ng, P.C., Henikoff, S.: Sift: predicting amino acid changes that affect protein function. Nucleic Acids Res. **31**(13), 3812–3814 (2003)
22. Pfister, T., Charles, J., Zisserman, A.: Flowing ConvNets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1913–1921 (2015)
23. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595 (2013)

24. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3487–3494 (2013)
25. Pishchulin, L., et al.: DeepCut: joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4929–4937 (2016)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017)
27. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: database and results. Image Vis. Comput. **47**, 3–18 (2016)
28. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397–403 (2013)
29. Sapp, B., Taskar, B.: MODEC: multimodal decomposable models for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3674–3681 (2013)
30. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483 (2013)
31. Tian, Y., Zitnick, C.L., Narasimhan, S.G.: Exploring the spatial hierarchy of mixture models for human pose estimation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 256–269. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_19
32. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems, pp. 1799–1807 (2014)
33. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
34. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: a recurrent process applied for end-to-end face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4177–4187 (2016)
35. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
36. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539 (2013)
37. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1385–1392. IEEE (2011)
38. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: a step towards the solution. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops, p. 2 (2017)
39. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)

40. 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
41. Zhu, S., Li, C., Change Loy, C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4998–5006 (2015)
42. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE (2012)
43. Zhu, X., Lei, Z., Li, S.Z., et al.: Face alignment in full pose range: a 3D totalsolution. IEEE Trans. Pattern Anal. Mach. Intell. (2017)
44. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)