Score-Guided Face Alignment Network Under Occlusions

Xiang Yan, Huabin Wang(✉), Qi Wang, Jinjie Song, and Liang Tao

Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230031, China
xiang199286@gmail.com, {wanghuabin,taoliang}@ahu.edu.cn,
1054588756@qq.com, 3470755438@qq.com

Abstract. Recent state-of-the-art landmark localization task are dominated by heatmap regression and fully convolutional network. In spite of its superior performance in face alignment, heatmap regression method has a few drawbacks in nature, such as do not follow shape constraint and sensitivity to partial occlusions. In this paper, we proposed a score-guided face alignment network that simultaneously outputs a heatmap and corresponding score map for each landmark. Rather than treating all predicted landmarks equally, a weight is assigned to each landmark based on the two relational maps. In this way, more reliable landmarks with strong local information are assigned large weights and the landmarks with small weights that may stay with occlusions can be inferred with the help of the reliable landmarks. Meanwhile, an exemplar-based shape dictionary is designed to take advantage of these landmarks with high score to infer the landmark with small score. The shape constraint is implicitly applied in this way. Thus our method demonstrates superior performance in detecting landmarks with extreme occlusions and improving overall performance. Experiment results on 300 W and COFW dataset show the effectiveness of the proposed method.

Keywords: Face alignment · Fully convolutional network · Occlusion

1 Introduction

Face alignment [5,25,40], aslo known as facial landmark detection, which aims to find the locations of a set of predefined facial landmarks (e.g., mouth, eyes, nose, cheek and so on) in a face image. It is a crucial pre-processing step for face recognition [16,26,27], expression recognition [3,13], face analysis [21] and so on. As a well established problem in computer vision, researchers have proposed many methods and made significant progress in face alignment. Recently, heatmap regression method [4,6,10] has shown superior performance on face alignment. However, Face alignment under occlusions still remains unsettled. Especially, when face images suffer from heavy occlusions, the performance of face alignment drops severely.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 195–206, 2018.
https://doi.org/10.1007/978-3-030-03338-5_17

To address face alignment under occlusions, several methods are proposed to tackle face alignment under partial occlusions. The method of [7] divides face into a 3×3 grid and only draw features from the $1/9$ of facial region to several separate regressors. The work in [29] proposes a robust cascaded regression framework to handle large facial pose and occlusion. The landmark locations and the landmark visibility probability are updated stage by stage. The method of [18] treat face alignment as an appearance-shape model problem. They learn two dictionaries which are relational, one for the appearance of human face and one for the facial shape. By the two relational dictionaries, the face appearance is employed to infer occlusion and suppress the influence of occluded landmarks. The work in [33] cascades several Deep Regression networks (DR) and De- corrupt Auto-encoders (DA) to explicitly handle partial occlusion problem. In contrast with previous methods that only predict occlusion, the proposed De-corrupt Auto-encoders can recover the occluded facial appearance. They divide the facial landmarks to seven components, each specific DA is able to recover the occluded appearance. Although these methods have shown superior performance in aligning occluded faces, they have limited scalability and robustness. First is the lack of large-scale ground truth occlusion annotation for images in the wild. The task of providing occlusion annotation is often time-consuming, involving a considerable amount of tedious manual work. Another challenge is in the inherent complex facial appearance. Generally, the performance of appearance-shape dictionary depends on whether the image patterns reside within the variations described by the face appearance dictionary. Therefore, it shows limited robustness in unconstrained environment where appearance variations are too wide and complicated. In addition, recovering the occluded appearance is not without diffculties.

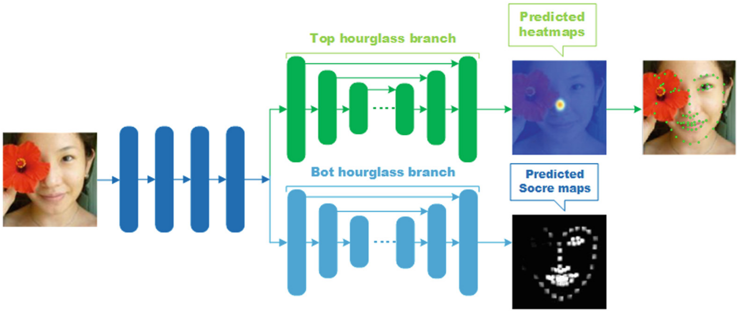


Fig. 1. Papers main idea: Given a face image as input, our network simultaneously outputs heatmaps and score maps. Due to part occlusions, the occluded landmark cannot be located precisely. Observe that the score for the occluded parts is much lower than that of the non-occluded parts in score maps. Based on the two relational maps, the occluded landmarks can be refined with the help of non-occluded landmark by exploiting geometric constraints of face shape.

In this paper, we propose a novel score-guided face alignment network to deal face with large occlusions. The key innovation of our method is score map which is able to dynamically select more reliable landmarks and use these reliable landmarks to refine the landmarks with small score. See Fig. 1 for a graphical representation of our paper's main idea. The proposed network outputs heatmaps and score maps. The occluded part is obvious in score map and has small score than non-occluded part. Rather than treat all landmarks equally, we assign a weight to each landmark based on heatmaps and score maps and the occluded landmark can be refined with the help of the non-occluded landmarks. More specifically, due to the partial occlusion, the occluded landmark cannot be located precisely. However, the non-occluded landmark can be located precisely. Since the non-occluded landmarks have lager weights than occluded landmarks. An exemplar-based shape dictionary act as shape priors can be utilized to search most similar shapes to reconstruct the face shapes based on the weights of landmarks.

The main contributions of our method can be summarized as follows:

1. We propose a novel face alignment network that simultaneously outputs heatmaps and score maps, which is more robust to occlusions. Note that no occlusion annotations are used.
2. Rather than treating all landmarks equally, we introduce score map to assign weight to each landmark. In this way, more reliable landmarks with large weights can help to refine the occluded landmarks with small weights.

2 Related Work

Prior to deep learning, cascade regression [9,17,18,22,23,37] is a popular method in face alignment, it starts with an initial facial shape and refine the shape in a cascaded manner. For each regressor, it learns a mapping function from shape-indexed features to the shape increment. The authors of [31] proposed a method named Supervised Descnet Method (SDM) to learn cascade regressors with strong handcrafted features such as SIFT. The work in [23] proposes learning local binary features by using random forests. Thanks to the sparse binary features, its speed can achieve 3000 FPS. To reduce the influence of inaccurate shape initializations, In [37] a coarse to fine search method is proposed. It begins with a coarse search over a shape pool and employs the coarse solution to finer search of shapes. The authors of [38] reformulates the popular cascaded regression scheme into a cascaded compositional learning (CCL) problem. It divides all training samples into several domains. Each domain-specific cascaded regressor handle one domain. The final shape is a composition of shape estimations across multiple predictions. The method of [11] trains multi-view cascaded regression models using a fuzzy membership weighting strategy, which improving the fault-tolerant of cascade regression. Although cascade regression has achieved good performances on the wild databases, inaccurate shape initializations, independent regressors and handcrafted features still may be sub-optimal for face alignment.

This conventional cascade regression, however, has been greatly reshaped by convolutional neural networks (ConvNets). Recent face alignment methods have universally adopted ConvNets as their main building block, largely replacing hand crafted features. The work in [36] uses multi-stage deep networks to detect facial landmarks in a coarse to fine manner. The authors of [35] formulates a novel tasks-constrained deep model to jointly optimize landmark detection together with the recognition of heterogeneous but subtly correlated facial attributes which improves the performance of landmark detection. The work in [34] employs Autoencoder netwroks (CFAN) that combined several stacked auto-encoder networks in a cascaded manner. The authors of [28] proposes a convolutional recurrent neural network architecture. The feature extraction stage is replaced with a convolutional network, the fitting stage is replaced with the Recurrent Neural NetWork. The work in [30] employs an Attention LSTM (A-LSTM) and an Refinement LSTM (R-LSTM), which sequentially selects the attention center by A-LSTM and refines the landmarks around the attention-center by R-LSTM. The authors of [19] presents a deep regression architecture with two stage reinitialization to explicitly deal with the initialization problem by face detection. FAN [6] employs stacked hourglass Network with a state-of-the-art residual block to solve the 2D&3D Face Alignment problem. The work in [10] formulate a novel Multi-view Hourglass Model which tries to jointly estimate both semi-frontal and profile facial landmarks.

3 Methodology

3.1 Network Architecture

Here, we describe our network architecture based on hourglass [20] backbone. The input is a face image with spatial resolution 128×128. The network starts a 7×7 convolutional layer with stride 2 and padding 3 to process the image to spatial resolution 64×64, followed by three residual blocks [14] to increase feature channels. Then the network is split in two sub-branches. The top sub-branch is a hourglass network, which is a symmetric top-down and bottom-up full convolutional network. Then two residual blocks process the feature maps to 128 channels. After that, nearest neighbor upsampling is used to increase the spatial resolution to 128×128, followed by a residual block and a convolutional layer with 1×1 kernels to produce heatmaps. The bottom sub-branch has the same network structure with the top sub-branch. Batch Normalization is used to before all convolutional layers expect the first convolutional layer with kernels 7×7. ReLU is the activation function. In summary, the input of network is a face image with spatial resolution 128×128. The network output N heatmaps and N score maps, where N is the number of landmarks. Each landmark corresponds to a heatmap and a score map (Fig. 2).

3.2 Score Map and Heatmap

Heatmaps are extensive used in landmark localization tasks. The model outputs N heatmaps where N is the number of landmarks. The pixel with the high- est

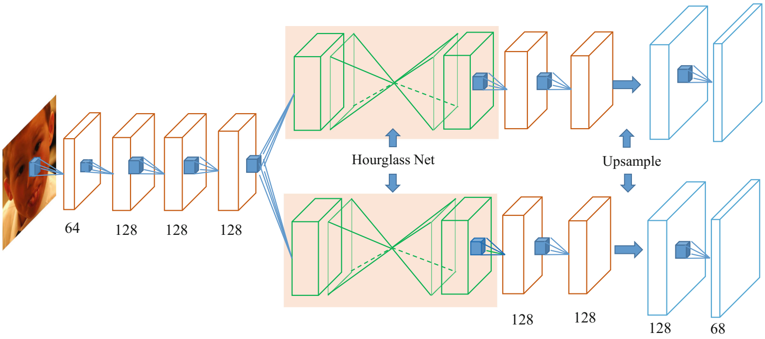


Fig. 2. An illustration of our network architecture.

value is used as the predicted landmark location. Great progress has been made by heatmaps. However, the landmarks with partial occlusion and complex background still cannot be precisely located. To deal with occlusions, we introduce score maps to assign weight to each landmark and suppress the influence of occlusions. During training, Heatmap for one landmark is created by putting a Gaussian peak at ground truth location of the landmark. While the score maps are binary maps, the values within a certain radius around the ground truth locations are set to 1 and the value for the remaining are set to 0. See Fig. 3 for example outputs produced by our network. The non-occluded face part has higher score than the occluded-part in score map. Rather than treating all landmarks equally, we weight each landmark based on their values in score maps. In this way, more reliable landmarks with strong local information are assigned high weights. The landmarks with small weights that may stay with occlusions can be refined with the help of reliable landmarks. Based on the two relational maps, the process of assigning weight can be written via the equation

$$w_i = \frac{\sum_{k=X_i-r}^{X_i+r} \sum_{t=Y_i-r}^{Y_i+r} score_i(k,t)}{(2*r+1)^2}. \tag{1}$$

where $score_i(k,t)$ is the value of coordinate (k,t) in i-th score map. X_i and Y_i are the predicted locations of i-th landmark.

3.3 Face Shape Reconstruction

Based on the two relational maps, the weight of each landmark can be determined. For the non-occluded face images, the heatmaps and score maps assign high weights to each landmark. The final predicted face shape is the locations decoded from heatmaps. For the heavy occluded face images, score maps only can check out these inaccurate landmarks with small weights, these landmarks still cannot be accurately located. Intuitively, the predicted face shape should

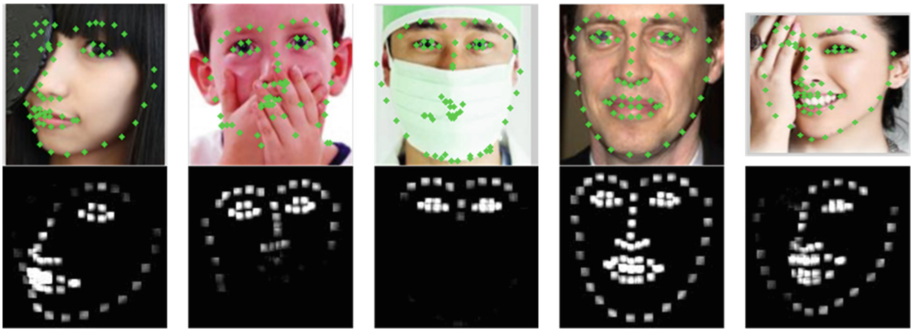


Fig. 3. Example outputs produced by our proposed network. First row shows landmark locations decoded by heatmaps. Second row shows the proposed score maps. Observe that the occluded landmarks cannot be precisely located in most cases. The non-occluded parts in score maps have higher score and are clearer than the occluded parts.

look like a face shape. Human vision has ability to predict good face shape by exploiting geometric constraints. Motivated by this, these inaccurate landmarks caused by occlusions can be refined by searching the most similar face shapes based on non-occluded landmarks, which is feasible and simple.

However, searching from all training samples is time-consuming. There are a lots of similar face shapes which are redundant. Assuming there are M training samples in train set. When M is large, searching from all training samples would be time-consuming. Follow [18], We apply K-SVD [1] on all training shapes to get N representative face shapes and use these face shapes as a shape dictionary D_S. Searching from D_S will be more effective. The searching process is formally written as

$$min_{s_1 \cdots s_k} \|W^S S - (W^S S \odot W^S D_S)\|_2^2 \tag{2}$$

where $W^S = diag(w_1, \cdots, w_N, w_1, \cdots, w_N)$ is the weight matrix and the w_i is the weight of the i-th landmark calculated via Eq. 1. The goal of W is to force the search process to emphasize on the landmarks with high weights and ignore the landmark with small weights. $s_1 \cdots s_k$ are the k nearest exemplar shapes of the non-occluded landmarks. After that, the occlusions landmarks can be reconstructed by the k nearest exemplar shapes and the reconstruction coefficients can be computed by least squares method (Fig. 4).

3.4 Training Details

During training, to prevent overfitting, all training samples are augmented by random in-plane rotation (from -30^o to $+30^o$), translation, scale (from 0.9 to 1.2), flip and adding color jittering. The network input is a RGB image of size 128×128. The network is optimized by RMSProp with an initial learning rate of 0.0001 and drop to 0.00005 after 20 epochs. All models are trained using

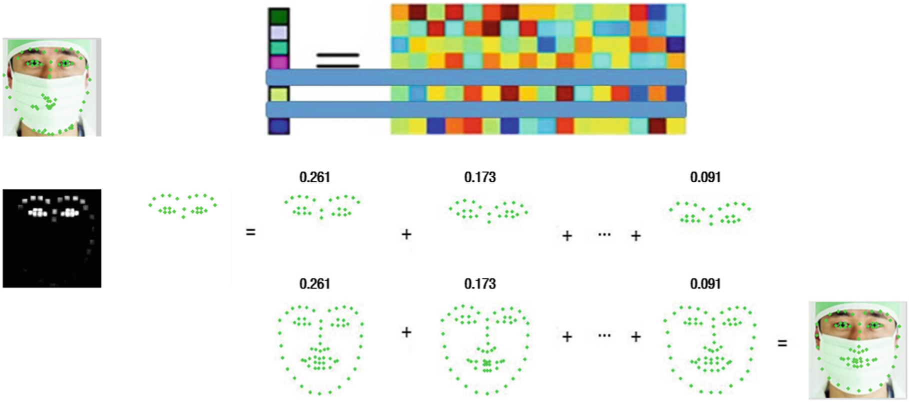


Fig. 4. Face shape reconstruction by the k nearest exemplar shapes.

PyTorch with a Nvidia 1080-Ti GPU card with a mini-batch size of 10 for 80 epochs. The loss function is defined as

$$Loss = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\|H_k - \hat{H}_k\|_2^2 + \frac{\lambda}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\|S_k - \hat{S}_k\|_2^2 \tag{3}$$

where N is the number of training samples,H_k and $\hat{H}_k$ are the predicted heatmaps and the ground-truth heatmaps. S_k and $\hat{S}_k$ are the predicted score map and ground-truth score map. λ is a hyperparameter to balance the loss functions. During inference, the predicted landmark locations $\hat{Z}_k$ is decoded from the predicted heatmap H_k by taking the locations with the maximum value as follows,

$$\hat{Z}_k = \arg\max_p H_k(p) \tag{4}$$

4 Experiments

4.1 Datasets

For training, 300-W is the most widely-used in-the-wild dataset for 2D face alignment. All face images are labeled by 68 landmarks. The training set consists of AFW [39] dataset, HELEN [15] training set and LFPW [2] training set, there are 3148 face images in total. For testing, we report the results on LFPW testing set, Helen testing set and IBUG dataset. To verify the effectiveness of our method on occluded faces, we evaluate COFW [7,12] testing set. The COFW dataset is a challenging dataset with severe facial occlusions and large facial pose collected from web. There are 1345 face images in training set and 507 face images in testing set. All face images are labeled by 29 landmarks. Since our model is trained on images with 68 landmarks, Follow [12], we use the COFW with 68 landmarks for testing. Note that we only use COFW testing set for evaluation.

4.2 Metrics

Given the predicted landmark locations and ground-truth landmark locations, the Normalized Mean Error (NME) or cumulative error distribution (CED) curves employed to evaluate the localization performance. The normalization is normalized by inter-pupil distance and the NME is computed as follows:

$$error = \frac{1}{M}\sum_{i=1}^{M}\frac{\frac{1}{N}\sum_{j=1}^{N}\|p_{i,j}^{pred} - p_{i,j}^{gt}\|_2}{\|p_{i,l} - p_{i,r}\|_2} \tag{5}$$

where M is the number of testing images, N is the number of landmarks. $p_{i,l}, p_{i,r}$ are the locations of left eye center and right eye center in i-th face image. $p_{i,j}^{pred}$ is the predicted location of landmark location of the j-th landmark in i-th face image. $p_{i,j}^{gt}$ is the ground-truth location of landmark location of the j-th landmark in i-th face image.

4.3 Evaluation Results on 300W

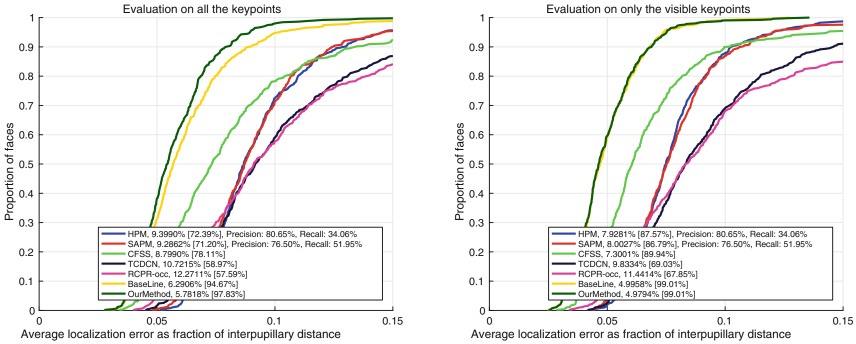
The 300-W [24] testing set consists of common set and challenging set. The com- mon set are Helen testing set and LFPW testing set. The challenging set is the IBUG dataset. Table 1 show the resluts on 300 W dataset. We compare our method with eleven state-of-the-art face alignment methods with RCPR [7], CFAN [34], ESR [8], SDM [31], LBF [22], CFSS [37], TCDCN [35], DNN [32], MD- M [28], RAR [30], TR-DRN [19]. Our method outperform most of these methods except RAR.

4.4 Evaluation Results on COFW

To verify the effectiveness of our method on various occluded face images, we test our method on COFW [7,12] dataset. The CED curves are shown in Fig. 5. It can be seen our baseline still outperform all other methods by a large margin. That is because our method benefits from heatmap regression and network architecture. By adding occlusion inference and face reconstruction, the NME error decreases from 6.29% to 5.78%. The success rate increases from 94.67% to 97.83%. Moreover, we analyse the evaluation on only the visible landmarks, our method and baseline show similar results on NME error and success rate. It can be concluded that heatmap regression method achieves excellent performance in detecting non-occluded face part. While evaluation on all the landmarks, benefit from score map to assign weight to each landmark and refine the occluded region by face reconstruction, our method show better results than baseline both in NME error and success rate.

Table 1. Landmark detection results on different subsets of the 300-W dataset in terms of the NME averaged over all the test samples.

Method	Common set	Challenging set	Full set
RCPR	6.18	17.26	8.35
SDM	5.57	15.40	7.52
ESR	5.28	17.00	7.58
CFAN	5.50	16.78	7.69
DeepReg	4.51	13.80	6.31
LBF	4.95	11.98	6.32
CFSS	4.73	9.98	5.76
TCDCN	4.80	8.60	5.54
DDN	-	-	5.59
MDM	4.83	10.14	5.88
RAR	4.12	8.35	4.94
TR-DRN	4.36	7.56	4.99
SIR	4.29	8.14	5.04
Ours	4.16	7.54	4.78



(a) (b)

Fig. 5. Comparison of different models on the COFW dataset: (a) evaluation on all the keypoints, (b) evaluation on only the visible keypoints.

5 Conclusion

In this paper, we propose a score-guided face alignment network which is robust to occlusions. The network simultaneously outputs a heatmap and corresponding score map for each landmark. Based on the two relational maps, more reliable landmark are assigned large weights and landmarks with small weights can be inferred with the help of the reliable landmarks. Experiment results on 300 W and COFW dataset show the effectiveness of the proposed method and showed significant performance improvements over the state-of-the-arts.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Sig. Process. **54**(11), 4311–4322 (2006)
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: Computer Vision and Pattern Recognition (CVPR), pp. 545–552 (2011)
3. Bettadapura, V.: Face expression recognition and analysis: the state of the art (2012). arXiv preprint arXiv:1203.6722
4. Bulat, A., Tzimiropoulos, G.: Convolutional aggregation of local evidence for large pose face alignment (2016)
5. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: The IEEE International Conference on Computer Vision (ICCV), vol. 1, p. 4 (2017)
6. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D and 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: International Conference on Computer Vision (ICCV), vol. 1, p. 4 (2017)
7. Burgos-Artizzu, X.P., Perona, P.: Robust face landmark estimation under occlusion. In: International Conference on Computer Vision (ICCV), pp. 1513–1520 (2013)
8. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014)
9. Deng, J., Liu, Q., Yang, J., Tao, D.: M3 CSR: Multi-view, multi-scale and multi-component cascade shape regression. Image Vis. Comput. **47**, 19–26 (2016)
10. Deng, J., Trigeorgis, G., Zhou, Y., Zafeiriou, S.: Joint multi-view face alignment in the wild (2017). arXiv preprint arXiv:1708.06023
11. Feng, Z.H., Kittler, J., Christmas, W., Huber, P., Wu, X.J.: Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3681–3690. IEEE (2017)
12. Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: detecting and localizing occluded faces (2015). arXiv preprint arXiv:1506.08347
13. Guo, Y., Zhao, G., Pietikäinen, M.: Dynamic facial expression recognition with atlas construction and sparse representation. IEEE Trans. Image Process. **25**(5), 1977–1992 (2016)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
15. Le, V., Brandt, J., Bourdev, L., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European Conference on Computer Vision (ECCV), pp. 679–692 (2012)
16. Li, D., Zhou, H., Lam, K.M.: High-resolution face verification using pore-scale facial features. IEEE Trans. Image Process. **24**(8), 2317–2327 (2015)
17. Liu, Q., Deng, J., Tao, D.: Dual sparse constrained cascade regression for robust face alignment. IEEE Trans. Image Process. **25**(2), 700–712 (2016)

18. Liu, Q., Deng, J., Yang, J., Liu, G., Tao, D.: Adaptive cascade regression model for robust face alignment. IEEE Trans. Image Process.(TIP) **26**(2), 797–807 (2017)
19. Lv, J.J., Shao, X., Xing, J., Cheng, C., Zhou, X., et al.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: Computer Vision and Pattern Recognition (CVPR), vol. 1, p. 4 (2017)
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. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), pp. 17–24. IEEE (2017)
22. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment via regressing local binary features. IEEE Trans. Image Process.(TIP) **25**(3), 1233 (2016)
23. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1692 (2014)
24. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: Conference on Computer Vision Workshops (CVPRW), pp. 397–403 (2014)
25. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Computer Vision and Pattern Recognition (CVPR), pp. 3476–3483 (2013)
26. Tai, Y., Yang, J., Zhang, Y., Luo, L., Qian, J., Chen, Y.: Face recognition with pose variations and misalignment via orthogonal procrustes regression. IEEE Trans. Image Process. **25**(6), 2673–2683 (2016)
27. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: DeepFace: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
28. 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: Computer Vision and Pattern Recognition (CVPR), pp. 4177–4187 (2016)
29. Wu, Y., Ji, Q.: Robust facial landmark detection under significant head poses and occlusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3658–3666 (2015)
30. Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.: Robust facial landmark detection via recurrent attentive-refinement networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 57–72. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_4
31. Xiong, X., Torre, F.D.L.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (2013)
32. Yu, X., Zhou, F., Chandraker, M.: Deep deformation network for object landmark localization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 52–70. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_4
33. Zhang, J., Kan, M., Shan, S., Chen, X.: Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders. In: Computer Vision and Pattern Recognition (CVPR), pp. 3428–3437 (2016)

34. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_1
35. 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
36. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 386–391 (2013)
37. Zhu, S., Li, C., Chen, C.L., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Computer Vision and Pattern Recognition (CVPR), pp. 4998–5006 (2015)
38. Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3409–3417 (2016)
39. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)
40. 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)