Facial Landmark Detection Under Large Pose

Yangyang Hao[1], Hengliang Zhu[1], Zhiwen Shao[1], Xin Tan[1], and Lizhuang Ma[1,2(✉)]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
{haoyangyang2014,hengliang_zhu,shaozhiwen,tanxin2017}@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn

[2] Department of Computer Science and Software Engineering, East China Normal University, Shanghai, China

Abstract. Facial landmark detection is a necessary step in many vision tasks and plenty of excellent methods have been proposed to solve this problem. However, for the conditions with large pose and complex expression, these works usually suffer an eclipse. In this paper, we propose a two-stage cascade regression framework using patch-difference features to overcome the above problem. In the first stage, by applying the patch-difference feature and augmenting the large pose samples to the classical shape regression model, salient landmarks (eye centers, nose, mouth corners) can be located precisely. In the second stage, by applying enhanced feature section constraint to the patch-difference feature, multi-landmark detection is achieved. Experimental results show that our algorithm has a significant improvement compared to the classical shape regression method and achieves superior results on COFW dataset.

Keywords: Facial landmark detection · Large pose
Patch-difference feature · Feature section constraint

1 Introduction

Facial analysis and processing technologies become hot topics in recent years. Facial landmark detection aims to find the feature points of organs (nose, eyes, mouth and cheek). This technique has extensive applications, such as face recognition [5], face tracking [21], facial beautification [6], expression recognition [14]. It is time consuming and inefficient to detect one landmark with a respective model and the most popular way is to treat all landmarks as a whole. Cascade shape regression model can efficiently regress all the landmarks at the same time and lots of approaches [4,7,11,15] based on this model have been proposed. However, as these methods hardly handle the scenarios of large pose and complex expression, the accuracy obviously decreases on dataset with these situations.

There are two main reasons for the above problem. One reason is that features are unstable and don't contain enough information. For instance, the famous

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11304, pp. 684–696, 2018.
https://doi.org/10.1007/978-3-030-04212-7_60

pixel-difference feature [4] which is the intension difference between two pixels. The pixel-difference feature plays an important role in the tree-based cascade shape regression methods. Though the feature is highly efficient, the pixel is sensitive to noise and too less information is used. Additionally, the pixel pairs are selected from a large number of candidates and it is a problem to select the most useful ones. In order to address the above problems, this paper proposes a patch-based feature to improve the performance of classical cascade regression methods. We use mean of image patch to replace the pixel to enhance the ability of noise immunity. The feature is normalized to make it scale-invariance at the same time. The new feature is more robust on variation of occlusion and illumination. In the procedure of selecting the best features from a large pool, we also propose a new feature selection constraint. Researchers [11] show that the closer between the pixel pairs, the features are better. We assume that each pixel has its nearest landmark and the new constraint: the distance between the pixel pairs should be smaller than the distance between their respective nearest landmarks. By combining these two constraints, the selected features are better.

Another reason for the above problem is that the datasets don't contain enough variations of pose and expression. For example, LFPW [2] (29 landmarks) is a dataset with little variation and LBF [15] can achieve the mean error of 3.35%. Another method cGPRT [13] reports a result of 4.63% on HELEN [12] (194 landmarks) dataset. However, the mean error on challenging set of 300W (68 landmarks) is around 10%. In first two data sets, most of the faces are natural and frontal, 300W challenging set contains many large pose faces that include in-plane and out-of-plane. In terms of the large roll angles, such as over 30°, testing data is much more than training data. We can see from Table 1, it is hard to get a good model on the data without enough pose variations. In this paper, we use a hard samples augmentation method to enrich the diversity of dataset. The idea of data augmentation is motivated by deep learning methods. The training data for deep learning methods are massive by leveraging preprocessing (translation, rotation, horizontal flip and compression), generally over half of a million. Aiming at large pose, we enlarge the training data through rotation and apple it to the cascade shape regression methods. However, the training capacity of conventional methods is from thousands to tens of thousands, only a small number of hard samples with large roll angles are selected for augmentation. In this way, training data in the same order can produce a better model and give better results.

In this paper, based on the patch-difference feature, we propose a two-stage cascade regression facial landmark detection method. We mainly cope with the scenario of large pose and improve the robustness of features for classical cascade regression. In summary, the contributions of this paper are:

- We propose a new patch-difference feature for tree-based cascade regression framework. By leveraging the patch information, our method is more accurate and has little affect to efficiency.
- We propose an enhanced feature selection constraint by using the information of the nearest landmarks of the feature pairs.

– A data augmentation method is used for face alignment that only a small number of large pose samples are augmented.

The remainder of this paper is organized as follows: Sect. 2 provides an overview of related work. Two-stage cascade regression method is presented in Sect. 3. Section 4 shows the experimental results and analysis. Section 5 is the conclusions.

2 Related Work

Facial landmark detection raises from last century and plenty of work have been proposed up to now. Generally, these approaches can be categorized into traditional methods and deep learning methods.

2.1 Traditional Methods

In recent years, the shape regression models [4,7,11,15,18,22] are extensively applied in face alignment. Cascade shape regression model is first used in [7] to estimate the facial shape and it is widely used in this field. ESR [4] directly learns a regression function to infer the shape from a sparse subset of pixel intensities indexed relative to the current shape estimate. Ensemble of Regression Trees (ERT) [11] substitutes the fern weak regressor in ESR [4] with a regression tree and limits the distance between the pairwise feature points to achieve a better result. Local Binary Feature (LBF) [15] proposes to learn local binary feature for each landmark independently and jointly regresses for all landmarks. Supervised descent method (SDM) [22] predicts shape increment by employing a cascaded linear regression based on SIFT features. GSDM [21] improve the performance of SDM [22] by computing the gradient in global. CFSS [26] applies the idea of coarse-to-fine to do shape searching in the sub-region and the results are not affected by the initial shape. Similar to ESR [4], LBF [15] and cGPRT [13], we focus on discriminative feature and propose a new feature to improve the performance.

2.2 Deep Learning Methods

Deep learning methods are the most popular in present. Sun *et al.* [16] first apply cascaded deep convolution network to estimate the position of five facial landmarks and refine the position of landmarks level by level. Zhou *et al.* [25] also use multi-level deep networks to detect facial landmarks in a coarse-to-fine manner. Honari *et al.* [9] present Recombinator Networks by using multi-scale input maps for learning coarse-to-fine feature. TCDCN [24] proposes a multi-task learning method that employs auxiliary facial attribute recognition to obtain correlative facial properties to improve the performance.

3 Two-Stage Cascade Regression

3.1 Overview of Our Method

Our method includes two main parts: salient landmark detection and multi-landmark detection. In the first stage, salient landmark detection is used to obtain positions of salient landmarks, then an initial face is generated as input for the next stage. Salient landmark set is the smallest subset that can roughly represent the characteristic of a face. Therefore, it is rational to leverage this information to generate the initial face for the next stage. The initial face is computed by a linear combination of several similar faces. Similar faces are obtained by searching from training samples according to distances between each other. In this paper, Manhattan distance of salient landmarks is applied. The weight w_n for each similar face is computed as follows:

$$w_n = \frac{\frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{N}}{N} \tag{1}$$

where $n = 1$ represents the most similar one, N is the number of similar faces and we use 19 similar faces, this formulation ensures that the more similar face has a bigger weight. In both of two stages, tree-based cascade regression framework is applied. Mean face and generated face are used as initialization in regression procedure for two stages respectively. Training data augmentation and patch-difference feature are used in the first stage to achieve the precise locations of salient landmarks. Enhanced feature selection constraint is applied to patch-difference feature in the second stage. Because salient landmarks are only 5 points, the distance between the nearest landmarks of pairwise points always bigger than the distance between pairwise points. That is to say, the new feature selection constraint is satisfied by default for salient landmark detection.

3.2 Tree Based Cascade Regression Model

A single regression model is insufficient for facial landmark detection in the wild that contains large variations of pose, expression, illumination and occlusion. Therefore, researchers tend to use a sequence of regressors to refine the results step by step. Tree model is generally used in training procedure of the regressors. Firstly, we give a brief introduction of cascade process. The shape of a face can be presented as $S = \{X_j | j = 1, 2...p\} \in \Re^{2p}$, p is the number of the landmarks, X_j denotes the x, y-coordinates of the j-th landmark in a face image I. By applying linear regression framework, formulation of cascade process is following: $S_{i,t+1} = S_{i,t} + r_t(I, S_{i,t})$, where r_t represents the t-th regressor, $S_{i,t}$ represents the current estimated shape of level t, $S_{i,t+1}$ represents the shape of next level. In this manner, the shape is updated step by step and increment for the shape of next level is r_t. And in each level, the regressor $r_t(I, S_{i,t})$ is learnt by solving the following optimization problem:

$$r_t = \arg\min_{r_t} \sum_{i=1}^{L} \|S^*_{i,t} - S_{i,t} - r_t\|_2 \tag{2}$$

where $S^*_{i,t}$ is the ground-truth, L represents the number of training data. Friedman [8] proposes gradient boosting tree algorithm to learn the regressor r_t and sum of square error loss is used in the algorithm. The number of levels is usually over 10. The idea of coarse-to-fine is exploited in the procedure.

Obviously, the crucial process is to learn a regressor r_t and we name it regression tree. The pixel difference feature is simple and geometric invariance in a certain intension, but it doesn't use the neighbor pixel information and is not a normalized feature. This paper addresses this problem and solves it in the later part. At each split node of the regression tree, threshold is applied to classify the samples into different leaf node according to the pairwise pixel difference value. Usually, at each node, we greedily select the best split from a number of candidate splits that are randomly generated. The best one should minimize the sum of the square error. Use θ to present the parameter set (τ, u and v), τ is a threshold, u and v are positions of pairwise points. This process can be represented in the following formulation:

$$E(M, \theta) = \sum_{s \in \{l,r\}} \sum_{i \in M_{\theta,s}} \|r_i - \mu_{\theta,s}\|^2 \tag{3}$$

$$\mu_{\theta,s} = \frac{1}{\|M_{\theta,s}\|} \sum_{i \in M_{\theta,s}} r_i \tag{4}$$

where M is the indices of training samples used in the node, $M_{\theta,l}$ is the set of indices of samples that are classified into the left node judged by the threshold, r_i is the residual of sample i in the gradient boosting algorithm. By omitting the parts that are independent of θ, the formulation above can be rewritten as follows:

$$\arg\max_{\theta} E(M, \theta) = \arg\min_{\theta} \sum_{s \in \{l,r\}} \|M_{\theta,s}\| {\mu_{\theta,s}}^T \mu_{\theta,s} \tag{5}$$

$\mu_{\theta,s}$ is the only factor that is to be computed and the node split optimization is efficient.

3.3 Patch-Difference Feature and Feature Selection

Pixel difference feature used in the regression tree is difference between intensities of two pixels in an image, it is highly efficient and accurate. ERT [11] can achieve 1000 fps (frame per second) for 68 landmarks detection. The pixel difference feature is simple and geometric invariance in a certain intension, but it is sensitive to noise. We propose a patch-difference feature to cope with this problem and try to use more potential information. Following is formulation used to compute the features:

$$\frac{MP(u) - MP(v)}{MP(u) + MP(v)} \tag{6}$$

where $MP(\cdot)$ is a function that computing the mean of an image patch, considering of the efficiency, we compute the mean of a 3×3 patch. In this way,

neighboring pixel information is used and the feature is a normalized form. This feature measures the relative difference between two image patches and the formulation in the same form as the Weber Fraction. The Weber's Law [10] is that the human perception of difference in stimulus is often measured as a fraction of the original stimulus. This form is robust against illumination changes. By leveraging the information of patches, this fraction form is robust to noise.

Candidates of features are generated randomly and this factor leads to a big difference between good feature and poor feature. On the other hand, the candidates pool should be big enough to make sure good features are contained. It is necessary to select a number of good ones from candidates pool and other work [11] has proposed a feasible measure. The constraint is that the pairwise points have a bigger probability to be selected when the distance between them is smaller. Exponential function is chosen to do this work, that is: $e^{-\lambda||u-v||}$, where $||\cdot||$ represents the Euclidean distance, λ is the parameter to control distance of the pairwise points. In this paper, we add an enhanced constraint that further improves the performance. We assume that each point corresponding to a nearest landmark. The additional constraint is that the distance between two landmarks should be bigger than the distance of pairwise points. The formulation is as follows,

$$||u - v|| < ||ul - vl|| \tag{7}$$

where ul and vl are the nearest landmarks of u and v, we show it in Fig. 1.

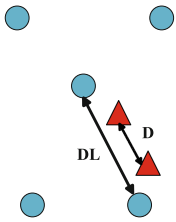


Fig. 1. Blue dots represent estimated landmarks of level t, red triangles represent two pixels. D is the distance between the pixel pair and DL is the distance between two landmarks.

3.4 Hard Sample Augmentation

For facial landmark detection, diversity of the annotated data sets is limited. The amount of the datasets from hundreds to thousands and most of the samples are frontal faces with natural expression. The performance decreases obviously when the faces have large variations in pose, expression, illumination and occlusion. 300W dataset is a good evidence to illustrate the above situation. This dataset includes two subsets: the common set and the challenging set. The mean error on common set is around 5% and the mean error on challenging set is around 10% for traditional methods.

Table 1. The analysis of face pose of 300 W dataset. The numbers represent quantity of samples with roll angle over 20°, 30° and 40°

300 W dataset	Training (3148)	Testing (689)
$> 20°$	83	34
$> 30°$	2	9
$> 40°$	0	2

The paper provides an analysis of face poses on the 300 W data set and roll angle of a face is regarded as the face pose. We use salient landmark information to analyze distribution of the roll angles. The roll angle is the angle between line LA and the Y-axis. Line LA is consisted of midpoint of eye centers and midpoint of mouth corners. The biggest roll angles of training and testing samples are 34° and 47°. From Table 1, we can see that the training data is seriously insufficient for large roll angles. The roll angle of training data is mainly below 30°, while some of testing samples with roll angle near 50°. Our method is following, first, the faces are classified into 5 categories (left, right, up, down and frontal) and 10 samples for each category are selected to be rotated ±30°, ±40° and ±50°. In a normalized face, the relative location of the nose tip is used to decide the category that a face belongs to. If nose tip at the left side of the center of the face, it is left face and 10 samples with the largest horizontal distance in this direction are selected. If nose tip on top of the center of the face, it is up face and 10 samples with largest vertical distance in this direction are selected. For frontal face, we choose 10 samples with the smallest Euclidean distance between nose tip and the center. Original 300 W data set is 3148 and the augmented data is 3448. The classified examples are showed in Fig. 2.

(a) (b) (c) (d) (e)

Fig. 2. These faces are samples of 5 different categories. The faces belong to up (a), down (b), left (c), right (d) and frontal (e).

4 Experiments

Datasets: Two challenging datasets are used for facial landmark detection to demonstrate our method achieves state-of-the-art performance. Faces of these datasets have a big variation on pose, expression, occlusion, and illumination.

300W dataset: It is a 68 landmarks dataset and consists of two subsets, the common subset and the challenging subset iBUG. Dataset configuration in [15]

is used to have a fair comparison. The training set contains 3148 images. Test set contains 689 images.

COFW dataset [3]: This dataset is annotated with 29 landmarks and mainly contains the faces with heavy occlusion, training samples are 1345 and 507 samples for testing.

Evaluation metric: Standard mean absolute error is used in the experiments. All errors are normalized by the inter-ocular distance and results in this section are simplified form without '%' symbol. For 300 W full set, Calculated Error Distribution (CED) curves are plotted to give more visible results.

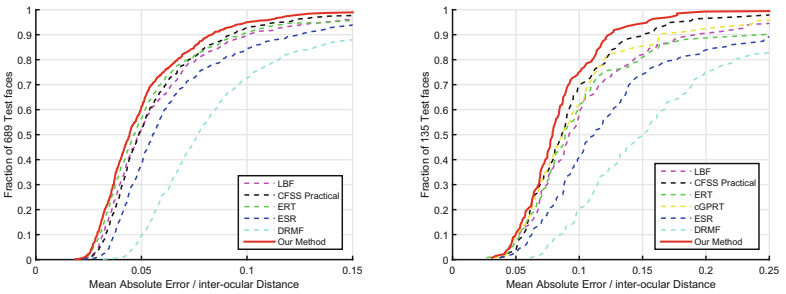
Parameter setting: In the training procedure, 20 randomly selected faces and 20 similar faces are used as initialization for salient landmark and multi-landmark detection respectively. Cascade level $T = 18$ and 15 are for first stage and second stage, $K = 500$ weak regressors form a strong regressor r_t, $D = 5$ is the depth of the regression tree. Shrinkage factor is 0.1. For node splitting, we repeat $S = 500$ times to find the best one. Following the feature selection constraint, 400 pairwise pixels and a randomly chosen threshold corresponding to each pair is used. The bounding boxes are provided in the database.

Table 2. Results of averaged error (%) are compared with state-of-the-art approaches on 300W. Errors are normalised by the inter-ocular distance, and the results of other methods are directly cited from the already published papers.

Method	Common	Challenging	Full set
DRMF [1]	6.65	19.79	9.22
ESR [4]	5.28	17.00	7.58
RCPR [3]	6.18	17.26	8.35
SDM [22]	5.57	15.40	7.50
ERT [11]	-	-	6.40
LBF [15]	4.95	11.98	6.32
cGPRT [13]	4.46	10.85	5.71
CFSS [26]	4.73	9.98	5.76
TCDCN [24]	4.80	8.60	5.54
MDM [17]	4.83	10.14	5.88
RDR [19]	5.03	8.95	5.80
RAR [20]	4.12	8.35	4.94
Our method	4.36	8.70	5.21

4.1 Comparison with Other Work

Table 2, a comparison with state-of-the-art methods is displayed. Compared methods include DRMF [1], ESR [4], RCPR [3], SDM [22], CFAN [23], ERT [11], LBF [15], cGPRT [13], CFSS [26], TCDCN [24], MDM [17], RDR [19] and RAR [20]. We can see that our method outperforms all the conditional methods and it is also comparable with deep learning method (TCDCN [24], MDM [17], RDR [19] and RAR [20]). 300W challenging subset mainly focuses on large pose and complex expression, with the help of hard sample augmentation, our method is robust adequate for variation of pose and expression.



(a) (b)

Fig. 3. Comparison of CED curves on 300 W full set and challenging set.

We provide comparison of CED curves with state-of-the-art approaches on 300 W full set (Fig. 3(a)) and challenging set (Fig. 3(b)). Compared methods include DRMF [1], ESR [4], ERT [11], LBF [15], CFSS Practical [26], and we can see that our approach better than others. The compared methods are re-implemented and some of the results are provided by authors. ESR [4] and ERT [11] are reproduced by ourselves with the error of 7.76 and 6.42. The result of LBF [15] is provided by the author, the codes of DRMF [1] and CFSS Practical [26] are available online. We also show some visible results of 300 W datasets in first two rows of Fig. 4. Though these faces with large variations in pose, our method achieves good performance by applying the proposed method.

Table 3. Results of averaged error (%) are compared with state-of-the-art approaches on COFW dataset.

Method	ESR [4]	RCPR [3]	SDM [22]	TCDCN [24]	RAR [20]	Our method
COFW	11.2	8.50	9.33	8.05	6.03	5.35

Comparison with state-of-the-art methods on COFW dataset is showed in Table 3. COFW dataset mainly focus on face with occlusion. This dataset is very challenge due to lots of faces with heavy occlusions. Compared methods include ESR [4], RCPR [3], SDM [22], TCDCN [24] and RAR [20]. From the Table, we can see that our method is much better than other methods. With the help of the salient-to-all manner, our method is robust under the conditions of occlusion. Last two rows faces of Fig. 4 are challenging samples of COFW dataset due to heavy occlusions. With the help of the proposed method, especially salient landmark detection, the effect of occlusion is declined.

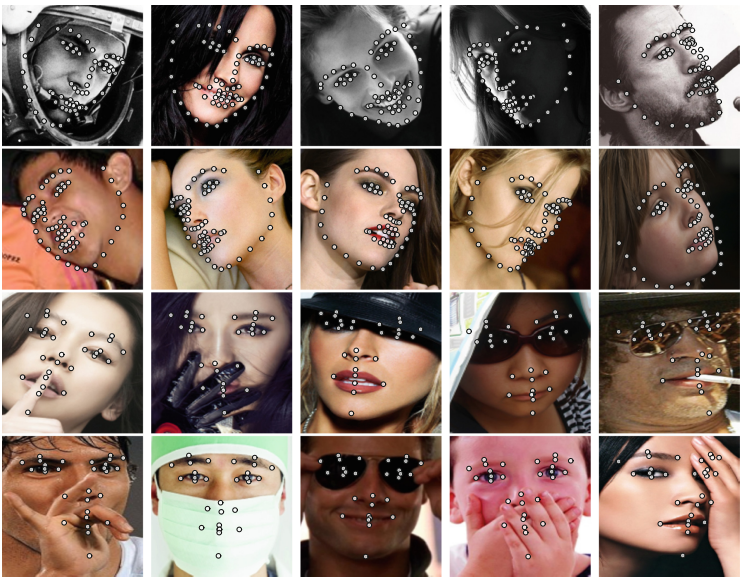


Fig. 4. Some challenging results of our method on 300 W and COFW dataset.

4.2 Incremental Analysis

In this paper, we propose three components to improve the performance and each of them shows a benefit to the whole process. Traditional cascade regression (CR) model is used as the baseline and three experiments are conducted to demonstrate effectiveness of our method. The three components are patch-difference feature (PD), enhanced feature selection constraint (FS), and hard sample augmentation (HSA). Both salient and 68 landmarks detection are conducted and two-stage is not applied. The new feature selection constraint is not applied in salient landmark detection because there are only 5 landmarks and this constraint is default satisfaction. Table 4 shows the performance of the three different components.

Table 4. Incremental analysis on 300W dataset.

Method	CR	CR+PD	CR+PD+FS	CR+PD+FS+HSA
Salient	4.39	4.23	-	3.95
Multiple	6.42	6.34	6.20	5.94

5 Conclusions

This paper presents a two-stage cascade regression framework. Salient landmark detection is done in the first stage and multi-landmark are detected in the next stage. Patch-difference feature, enhanced feature selection constraint and hard samples augmentation are applied in our algorithm. By utilizing the patch-difference feature and feature selection constraint, the feature maintains efficient and contains more information. With the augmentation, our method has a strong power to handle the condition with large pose. The performance improves significantly and increased training data is small compare to the original data. Our experiments are conducted on a single core Intel(R) Xeon(R) CPU E5-2630 v3 @2.4 GHz and speed is 190 fps. Experiments on two challenging data sets demonstrate the efficiency of our algorithm.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61472245) and the Science and Technology Commission of Shanghai Municipality Program (No. 16511101300).

References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Computer Vision and Pattern Recognition, pp. 3444–3451. IEEE, New York (2013)
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 545–552 (2013)
3. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: International Conference on Computer Vision, pp. 1513–1520. IEEE, New York (2013)
4. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vision **107**(2), 117–190 (2012)
5. Chen, C., Dantcheva, A., Ross, A.: Automatic facial makeup detection with application in face recognition. In: International Conference on Biometrics, pp. 1–8. IEEE, New York (2013)
6. Guo, D., Sim, T.: Digital face makeup by example. In: Computer Vision and Pattern Recognition, pp. 73–79. IEEE, New York (2009)
7. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: Computer Vision and Pattern Recognition, pp. 1078–1085. IEEE, New York (2010)
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232 (2001)

9. Honari, S., Yosinski, J., Vincent, P., Pal, C.: Recombinator networks: learning coarse-to-fine feature aggregation. In: Computer Vision and Pattern Recognition, pp. 5743–5752. IEEE Computer Society, New York (2016)
10. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall (1989)
11. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Computer Vision and Pattern Recognition, pp. 1867–1874. IEEE, New York (2014)
12. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_49
13. Lee, D., Park, H., Yoo, C.: Face alignment using cascade Gaussian process regression trees. In: Computer Vision and Pattern Recognition, pp. 4204–4212. IEEE, New York (2015)
14. Ramirez Rivera, A., Castillo, R., Chae, O.: Local directional number pattern for face analysis: face and expression recognition. IEEE Trans. Image Process. **22**(5), 1740–1752 (2013)
15. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Computer Vision and Pattern Recognition, pp. 1685–1692. IEEE, New York (2014)
16. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Computer Vision and Pattern Recognition, pp. 3476–3483. IEEE Computer Society, New York (2013)
17. 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, pp. 4177–4187. IEEE, New York (2016)
18. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: Computer Vision and Pattern Recognition. IEEE, New York (2015)
19. Xiao, S., et al.: Recurrent 3D–2D dual learning for large-pose facial landmark detection. In: IEEE International Conference on Computer Vision, pp. 1642–1651. IEEE Computer Society, New York (2017)
20. 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
21. Xiong, X., De la Torre, F.: Global supervised descent method. In: Computer Vision and Pattern Recognition, pp. 2664–2673. IEEE Computer Society, New York (2015)
22. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition, pp. 532–539. IEEE, New York (2013)
23. 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

24. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. **38**(5), 918–930 (2016)
25. Zhou, E., Fan, H., Cao, Z., Jiang, Y.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: International Conference on Computer Vision Workshops, pp. 386–391. IEEE Computer Society, New York (2014)
26. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Computer Vision and Pattern Recognition. IEEE, New York (2015)