IET Research Journals

Journals
The Institution of Engineering and Technology

Feedback Cascade Regression Model for Face Alignment

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Yangyang Hao[1*] *Hengliang Zhu*[1] *Zhiwen Shao*[1] *Lizhuang Ma*[1,2*]

[1] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China*
[2] *Department of Computer Science and Software Engineering, East China Normal University, Shanghai, China*
* *E-mail: haoyangyang2014@sjtu.edu.cn,ma-lz@cs.sjtu.edu.cn*

Abstract: Face alignment has made a great progress in recent years and cascade regression framework is one of the main contributors. However, performance of this framework is unsatisfied on the faces of heavy occlusion and large pose. Blame on that regression is sensitive to invisible landmarks and unified initialization is easy to make the results trapping into local minima. In this paper, we propose a new pipeline of salient-to-inner-to-all to progressively compute the locations of landmarks. Additionally, a feedback process is utilized to improve the robustness of regression. We bring out a pose-invariant shape retrieval method to generate the discriminative initialization. Experiments are performed on two benchmarks, and the experimental results demonstrate that the proposed method has a considerable improvement on cascade regression model which can achieve favorable results comparing with the state-of-the-art deep learning based methods.

1 Introduction

Social media data about faces has an explosive increasement in the last decade. Tons of face analysis related technologies have been developed. Face alignment is a necessary step in the whole face analysis pipeline, such as face recognition [1, 2], face tracking [3], facial beautification [4, 5], age estimation [6] and expression recognition [7, 8]. Face alignment aims to locate semantic facial landmarks (on eyes, nose, mouth and cheeks), some of the researchers address this problem by processing all the landmarks together while other researchers locate the landmarks individually. Cascade shape regression is a popular framework that operates all landmarks together. It can produce good results in an exact and fast way. But some landmarks are invisible which caused by the partial occlusion and pose variations, in this case, most of cascade shape regression based methods will give the incorrect results.

One reason leading to the above problem is the limitation of the regression framework. Popular cascade regression models [9–11] treat all landmarks as a indivisible whole and regress them together. However, some landmarks are invisible in the conditions of heavy occlusion, bad illumination or large pose. It is difficult to directly locate all landmarks of the above conditions, because learning the detection of the invisible landmarks is unpredictable. Another reason is that the initialization for regression is unified. Therefore, regression models start from an averaged initial shape will regress to different results. The averaged face is a natural frontal face. The result will be more accurate if the input face is similar to average face. It is not reasonable to regress a large pose face from an averaged initial face.

In this paper, we propose a salient-to-inner-to-all framework combined with a feedback operation to address the first problem. Firstly, 5 salient landmarks (eyes centers, nose tip, mouth corners) are located by the coarse cascade shape regression model. Then, inner landmarks that related to salient landmarks are located in the fine cascade shape regression. Inner landmarks are feature points that do not contain points of cheeks, such as the third column image of Fig. 1. Thirdly, all landmarks results are regressed. We use results of all landmarks to compute the new salient landmarks, then we replace the salient landmarks of first step by using new salient landmarks. Finally, inner landmarks are located again by applying the new salient landmarks information and the final all landmarks results are obtained. Salient landmarks detection can efficiently reduce the effect of the problem that some landmarks are invisible. It is possible to get all landmarks by directly utilizing 5 salient landmarks information. However, salient landmarks contain less structure information. Inner landmarks related to salient landmarks contain more information and the localization is more precise. Inner landmarks detection uses the local information and all landmarks detection uses the global structure information. Inner landmarks detection and all landmarks detection further improve the accuracy of salient landmarks. Therefore, it can ameliorate the salient landmarks with the results of all landmarks. Some bad cases of salient results can be corrected by the feedback. Accuracy of salient landmarks is improved after feedback operation and later stages can obtain a better benefit.

Initial shape is very important and more precise initial shape can lead to a more accurate result. For the second problem, we use a pose-invariant shape retrieval approach to help us to generate the discriminative initial shape. We assume that the initial face can be treated as the linear combination of some similar faces [12]. These similar faces can be searched from the training set. Pose-invariant shape retrieval method makes faces rotating into a uniform pose and Manhattan distance between landmarks of different faces is used to measure the similarity. Faces with high similarity are selected from training dataset to generate the initial shape. Although there are few faces with extreme variation in pose, some input faces with extreme pose variations can find enough similar faces for reference. The generated face is similar with the input face and it is discriminative according to the input face. By this way, the unified initialization is replaced by the discriminative initialization.

In this paper, we propose a feedback cascade shape regression framework (see Fig. 1) in which landmarks are progressively located in a salient-to-inner-to-all manner. The contributions of this paper are as follows: i) We propose a novel pipeline that is salient-to-inner-to-all way. Salient landmarks detection is robust in the conditions of heavy occlusion and large pose. Inner landmarks detection is an intermediate step to connect salient landmarks detection and all landmarks detection. ii) We utilize feedback scheme to ameliorate the classical cascade shape regression model. In this framework, results of salient landmarks detection are refined after feedback operation. iii) A pose-invariant shape retrieval method is proposed to search the similar faces to generate the discriminative initialization. By applying this method, the discriminative initial shape for the large pose input is more accurate. iv) Based on the new framework, conducted experiments show that our approach can improve the classical regression framework significantly. Extensive results show that our method is competitive with deep learning methods on 300W dataset. Our method is much better than all of other methods on COFW dataset and shows its robustness in condition of occlusion.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work. Feedback cascade shape regression framework and salient-to-inner-to-all manner are presented in Section 3. Section 4 shows the experimental results and analysis. Section 5 is the conclusions.

2 Related Work

In the last two decades, face alignment has a notable progress and lots of excellent work are reported. Generally, these approaches can be categorized into three classes, holistic methods, deep learning methods and local methods. The conventional way of holistic methods is that the model is trained by using all the landmarks together. Local methods train the model for a single landmark independently as the opposite. Deep learning methods use massive training data and different network architecture. In this paper, local methods and deep learning methods are introduced briefly and a large amount of space is used to introduce holistic methods.

2.1 Local Methods

In local methods, each landmark corresponding to a model, the model may be a detector, regressor or part template. Active Shape Model (ASM) [13] is the first work on this topic and learns patterns of variability from a training set to form the model. By using both shape and texture information, Constrained Local Model (CLM) [14] generates a set of regional template detectors and uses the generated response images to find the best matched image. Wang [15] proposes an algorithm to optimize the global warp update across all local search responses by enforcing convexity at each local patch response surface. Lucey [16] improves performance upon the canonical CLM formulation by applying linear SVMs as patch-experts and a composite warp update step. Saragih [17] proposed a new approximation of the likelihood maps by using nonparametric representations for the fitting procedure. Zhu [18] applies tree-structured models and uses global mixtures to capture topological changes corresponding to the viewpoint to obtain the results. Some researchers combine local and holistic methods to solve this problem. For example, [19] learns a model of the geometric relation between different face parts and integrates this part-based model into regression framework.

2.2 Holistic Methods

Active Appearance Model (AAM) [20] is a famous method in early time. AAM constructs face prior model by analyzing training data statistically and uses the model to match face images in the testing stage. In order to improve AAM fitting performance, Matthews [21] proposes an efficient fitting algorithm based on the inverse composition. Considering efficiency and accuracy, [22] uses a set of classifiers that learned from local patch to guide the search at the component level. In recent 10 years, shape regression model [9–11, 23–26] becomes one of the most classic frameworks and cascade shape regression is the most successful and widely used method. Cascade regression model is first used in [23] to estimate the facial shape. ESR [9] 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 weak fern regressor in ESR [9] with a regression tree and limits the distance between the pairwise feature points to achieve a better result. Local Binary Feature (LBF) [25] proposes learning local binary feature for each landmark independently and jointly regression for all landmarks. Supervised descent method (SDM) [10] predicts shape increment by employing a cascaded linear regression based on SIFT features. GSDM [3] improves the performance of SDM [10] by computing the gradient in global. cGPRT [27] applies Gaussian process into cascade regression trees and shape-indexed features to achieve good performance. CFSS [28] 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. In summary, Supervised descent method (SDM) [10], GSDM [3] and Project-out Cascade Regression (PO-CR) [26] focus on optimization problem, Explicit Shape Regression (ESR) [9], Local Binary Feature (LBF) [25]and cGPRT [27] focus on discriminative feature used in trees, CFAN [29] and CFSS [28] follow a coarse-to-fine manner. Xiao $et\ al.$ [30] propose a similar work that gradually increases the landmark that starts from 5 points to 19 points and finally extends to 68 points for face alignment. However, this method uses multi-initialization between the stages and these initializations come from K-means centers, as a contrast, we only use one initialization and it is generated by using training data. Our inner landmarks are 49 points that are highly associated with the 5 salient landmarks and do not contain landmarks on cheeks. 19 points of this method contain landmarks of cheeks and it is difficult to get accurate results. Because 19 points contain points of eyes, eyebrows, nose, mouth and cheeks, the structure of 19 points is a global structure. 49 points only contain points of eyes, eyebrows, nose and mouth, this structure is a local structure. Our salient landmarks are refined both in local and global. While this method refines the salient landmarks only in global, that is to say our framework is more comprehensive. CFSS [28] uses the idea of coarse-to-fine, but they use all landmarks in the whole procedure while we use salient landmarks firstly. They do shape searching for classification and we use shape searching to generate the discriminative initialization. Additionally, our shape searching method is much faster.

2.3 Deep Learning Methods

Deep learning based methods are the most popular in present and many deep learning approaches give a better results comparing to the traditional methods. Sun $et\ al.$ [31] first use cascaded deep convolution network to estimate the position of five facial landmarks and refine the position of landmarks level by level. Zhou $et\ al.$ [32] also use multi-level deep networks to detect facial landmarks in a coarse to fine manner. Honari $et\ al.$ [33] present Recombinator Networks by using multi-scale input maps for learning coarse-to-fine feature. TCDCN [34] proposes a multi-task learning method that employs auxiliary facial attribute recognition to obtain correlative facial properties to improve the performance of landmark detection.

3 Feedback Cascade Shape Regression

3.1 Cascade Shape Regression Model

Cascade shape regression model utilizes lots of regressors to make the initial shape regressing to ground truth. For facial landmark detection, the increments are offset of each landmark and the results are the location of landmarks. A face can be represented as $S = \{X_j | j = 1, 2...p\} \in \Re^{2p}$, where p is the number of the landmarks, X_j denotes the *x,y*-coordinates of the *j*-th landmark in a face image I. The cascade procedure is a linear process and the formulation of this process can be presented as follows:

$$S_{i,t+1} = S_{i,t} + r_t(I, S_{i,t}) \tag{1}$$

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 the next level. In this manner, the shape is updated step by step and increased the difference from the 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:

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

where S_i^* is the ground-truth, L represents the number of training data. Algorithm 1 shows the above procedure.

Gradient boosting tree algorithm [35] is applied and sum of square error is regarded as the loss to learn the regressor r_t. Offset of each

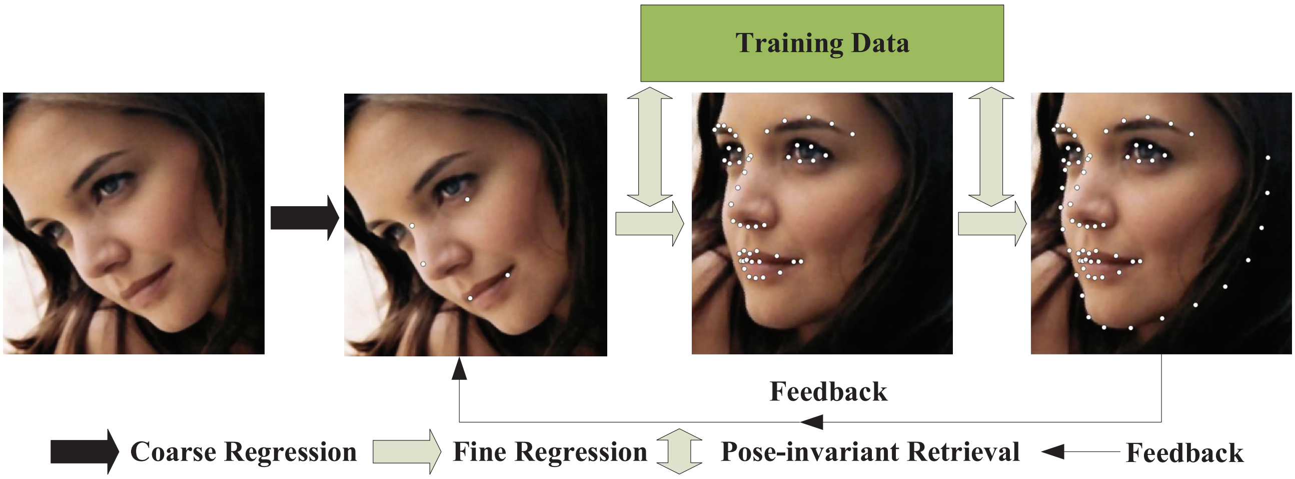

Fig. 1: The overall procedure of salient-to-inner-to-all face alignment. For an input face image, coarse regression is applied and salient landmarks are located. Then, a pose-invariant shape searching method is used to generate an initial shape for fine regression in inner landmark detection. All landmarks are detected like the former step and results are feedback to improve the accuracy of salient landmarks. Inner landmarks and all landmarks are located one more time.

Algorithm 1 Learning r_t in training stage.

```
Input: Training dataset $\{S_1, S_2, ..., S_L\}$;
1: Initialisation: $I_i, S_i^*, S_i$
2: For level $t$ from 1 to $T$
3: $\Delta S_{i,t} = S_i^* - S_{i,t}$, //Incremental of each level
4: $f_0(I, S_{i,t}) = \arg\min_{\gamma \in \Re^{2p}} \sum_{i=1}^{L} \|\Delta S_{i,t} - \gamma\|_2$
5: For weak regressor $k$ from 1 to $K$
       1) For $i$ from 1 to $L$
       $r_k(i) = \Delta S_{i,t} - f_{k-1}(I_i, S_{i,t})$
       2) By using a weak regression function $g_k$ to iteratively
   reach the targets $r_k(i)$.
       3) $f_k(I, S_t) = f_{k-1}(I, S_t) + v * g_k(I, S_t)$
Output: $\{r_t = f_{k=1}^{K}\}_{t=1}^{T}$
```

landmark is computed by averaging the samples belonging to the corresponding leaf node. At each split node of the tree, threshold is applied to classify the samples into different leaf node referring to the pairwise pixel difference value. Usually, at each node, we greedily select the best split from a number of candidates 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 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 this 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 residue of sample i in the gradient boosting algorithm. The formulation above can be rewritten as follows by omitting the parts that are independent of θ:

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

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

Cascade shape regression is successful in face alignment and many improvement methods are proposed. For example, [36] proposed a shrinkage factor 0< *v* <1 to control the increment. Because the number of levels is usually over 10, the increment should be smaller in later level to make sure the precision of regression. It is a very important factor to overcome the over-fitting. [11] proposed an effective constraint to help algorithm to select better features from a large random feature pool. They proposed the idea that the closer between the pairwise points in a face, the greater chose probability. In a big candidate feature pool, the distance is computed by the formulation: $e^{-\lambda||u-v||}$, where $|| \cdot ||$ represents the Euclidean distance, λ is the parameter to control distance of the pairwise points.

We use decision tree to learn the regressor r_t and local feature to make the decision for each tree. In this paper, a normalized feature called NPD [37] feature is used. Compared to the famous pixel difference feature, this feature is more robust and efficient. The mentioned two key technologies are also used in our framework.

3.2 *Feedback Regression in A Salient-to-inner-to-all Manner*

Cascade shape regression model starts from an initial shape. The face shape is updated through regressors in a sequence way. Our method applies this framework in a progressive and feedback manner. Salient landmarks are detected firstly, then inner landmarks which have a strong relationship with salient landmarks are located, all landmarks are detected after that. The positions of salient landmarks are updated with the results of all landmarks. Inner landmarks and all landmarks are regressed again to get the finally results. For salient landmarks detection, the mean shape is used as the initial shape and we call this regression " coarse regression ". For later stages, the discriminative initial shape is applied and we call this regression " fine regression ". The detail of generating the discriminative initial shape and pose-invariant shape retrieval method is presented in Section 3.3. The details of the testing procedure are described in Algorithm 2.

3.2.1 Salient-to-inner-to-all Manner: Heavy occlusion, large pose and bad illumination are the major problems for unconstrained face alignment. Some landmarks are invisible due to the above problems. The regression model hardly learns useful information from the missing points. To cope with the above problem, we obtain the results in a salient-to-inner-to-all way. Salient landmarks are located after coarse regression in the first step. This step provides a guide to the later steps. Due to sparsity and constitutive property, salient landmarks are insensitive to occlusion and pose variance. The salient points are used to search similar faces on the training set. We use

Algorithm 2 Overview of testing procedure.

```
Input: Training dataset, testing dataset and mean shape of 5 salient
    landmarks;
1: After coarse regression, salient landmarks $F_5$ is located,
2: Searching similar shapes $F_{sim5}$ by using Manhattan distance:
    $M(F_5^* - F_5)$, $F_5^*$ is ground truth salient point of training data,
3: Generating inner (49) landmarks initial shape $I_{49}$ by linear
    combination of $F_{sim5}$,
4: Fine regression by using $I_{49}$, we get inner landmarks $F_{49}$,
5: Searching similar shapes $F_{sim49}$ by using Manhattan distance:
    $M(F_{49}^* - F_{49})$, $F_{49}^*$ is ground truth 49 point of training data,
6: Generating initial shape $I_{68}$ by linear combination of $F_{sim49}$,
7: Fine regression by using $I_{68}$, we get all landmarks $F_{68}$,
8: Updating $F_5$ by using $I_{68}$, $F_5 \to F_{5new}$,
9: Repeating the step 2 to step 7 for $N_{re}$ time,
Output: Final results $F_{68new}$.
```

these similar faces to generate an initial shape for inner landmarks detection. The details of generating the initial face are presented in the next section. With the generated initial face, inner landmarks are located. Then, inner landmarks results are used to find similar faces and generate initial face for all landmarks regression. The generated initial face is similar to the input face and it is easy to be regressed to the target location.

It is not suitable to regress all landmarks directly from salient landmarks results, because the structure information of salient landmarks is not sufficient to generate a very discriminative initialization. For instance, salient landmarks can not distinguish faces with open mouth and close mouth. Different eyes and mouth expressions are treated as one kind. we can see this problem in Fig. 2. These two different faces have same salient landmarks and they can not be distinguished by only using salient landmarks. That is why we utilize inner landmarks as an intermediate step. Inner landmarks can help us to find more accurate similar faces and lead to a better initialization for all landmarks regression. These inner landmarks have a strong relationship with salient landmarks, so this intermediary step (inner landmarks detection) can keep advantages of salient landmarks detection.

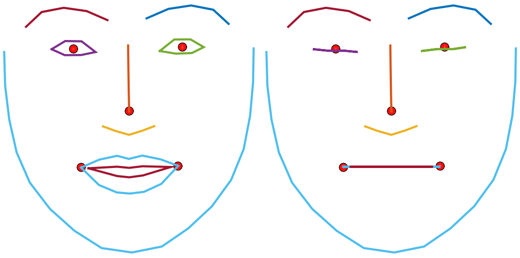

Fig. 2: Different faces with same salient landmark. Red points are salient points and different color lines are contour lines of face organs.

3.2.2 Feedback Regression: One of the limitations of cascade regression models is the initial unified shape. If the results trapped into the local optimum, the cascade regression procedure cannot jump out to find a global optimum. This drawback can be weakened by giving a discriminative initialization to some extent. In this section, we apply another strategy, that is feedback. Though results of all landmarks trapped into local optimum, salient points may be close to global optimum. Because salient landmarks are strong semantic and robust comparing to other landmarks. The salient landmarks positions are updated after all landmarks detection. In practical, we simply replace coarse result of the salient landmarks with the salient landmarks computed by using all landmarks. By this way, a new start is given and it is close to the global optimum. This procedure is illustrated in Fig. 3. The Fig. 3 (b) is the result which trapped into local optimum without feedback. From this figure, we can see that salient landmarks are close to ground truth even if all landmarks are not accurate. The Fig. 3 (c) and Fig. 3 (d) are updated salient landmark and final result after feedback. Obviously, feedback operation can improve the accuracy of salient landmarks when the coarse salient landmarks detection is not accurate and further improve the final result.

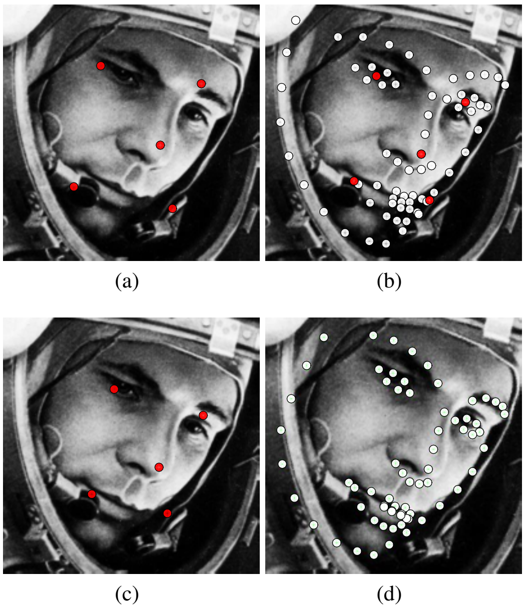

Fig. 3: Example of feedback. Image(a) is coarse result of salient landmarks, image(b) is result after using image(a). Image(c) is feedback result of salient landmarks by using image(b), image(d) is final result.

3.3 *Pose-invariant Shape Retrieval*

One of the important factors of regression is initialisation. Coarse regression is initialised with averaged shape and fine regression is initialise with generated shape. We assume that a face can be presented by a combination of similar faces, and weight of each face is proportional to the similarity. Based on the above assumption, the initial face can be obtained after we find some similar faces. Manhattan distance between two faces is calculated to measure the similarity. If the distance between two faces is small, the similarity will be high and these two faces are considered to be similar, vice versa. Weight w_n for each similar face is computed as follows:

$$w_n = \frac{\frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{N}}{N} \quad (6)$$

where $n = 1$ represents the most similar one, N is the number of similar faces. In our experiment, $N = 19$ can produce best result. This formulation makes sure that the more similar face has a bigger weight.

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

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

However, some of the samples with extreme pose cannot find enough faces with high similarity. From Table 1, we can see that samples with large pose is not enough if we want to find 19 similar faces. In this paper, we propose a pose-invariant shape retrieval method to find the similar faces. This method can reduce the incidence of this problem to some extent. All the training samples are rotated into a uniform pose that middle point between eye centers has the same x coordinate with the middle point of mouth corners.

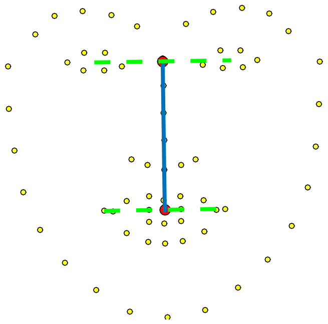

Fig. 4: The unified face. The blue line is consisted of two middle points.

Fig. 4 shows the unified face and we can see that the blue line should be vertical. A rotation matrix A is calculated to rotate the input face to the identical pose. We resize the rotated face into a fixed size, such as 100 × 100. The matrix A is obtained as follows:

$$A = \begin{bmatrix} cos(\theta_a) & sin(\theta_a) & 0 \\ -sin(\theta_a) & cos(\theta_a) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

where θ_a is the angle between the line consisted of two middle points and the vertical coordinate axis. Salient landmarks distance between two faces is used to find similar faces and then generate the initial shape for inner landmark detection. Inner landmarks distance between two faces is used to search similar faces and generating the initial shape for all landmark detection. The parameter θ_a can be computed easily after obtaining the salient landmarks and inner landmarks.

4 Experiments

Datasets: Though excellent performance has been reported on some datasets with little variations, it is still challenging to have a good result on other datasets with heavy occlusion and extreme pose. In this paper, experiments are conducted on two challenging datasets (300W and COFW) and state-of-the-art performance is presented .

300W dataset: This dataset is a 68 landmarks dataset and consists of five databases: AFW [18], LFPW [38], HELEN [39], XM2VTS [40] and challenging iBUG. Dataset configuration in [25] is used to have a fair comparison. The training set contains 3148 images. Test set contains 689 images. The dataset includes two subsets: the common subset and the challenging subset iBUG. The challenging subset contains some faces with large pose both in-plane and out-of-plane.

COFW dataset [41]: This dataset is annotated with 29 landmarks and mainly contains the faces with heavy occlusion. Number of training samples is 1345 and 507 samples for testing.

Evaluation Metric: Two metrics: standard mean absolute error and global mean absolute error [42] are used in the experiments. The commonly used evaluation metric: MAE (mean absolute error) is point-to-point distance between two faces. The GMAE (global mean absolute error) not only contains the point-to-point distance MAE, but also contains the structure distance D_G between two shapes. This evaluation metric is more comprehensive. All errors are normalized by the inter-ocular distance and results in this section are simplified form without '%' symbol. For 300W dataset, calculated error distribution (CED) curves and global CED curves are plotted to give more visible results. The standard mean absolute error is computed as follows:

$$MAE = \frac{||S - S^*||}{D_{in}} \tag{8}$$

where D_{in} is the inter-ocular distance, S is estimated shape and S^* is ground truth. The structure distance D_G between S and S^* is computed as follows:

$$D_G = \frac{\sum_{i=1}^{p} D(X_i, L_i^*)}{p * D_{in}} \tag{9}$$

$$D(X_i,L_i^*) = \frac{|(x_i^* - x_i)(y_{i+1}^* - y_i) - (x_{i+1}^* - x_i)(y_i^* - y_i)|}{||X_i^*(x_i^*, y_i^*) - X_{i+1}^*(x_i^*, y_i^*)||} \tag{10}$$

where p is the amount of points, $X_i(x_i, y_i)$ is landmark of estimated face S, $X_i^*(x_i^*, y_i^*)$ and $X_{i+1}^*(x_i^*, y_i^*)$ are landmarks of ground truth S^*, L_i^* is line consist of $X_i^*(x_i^*, y_i^*)$ and $X_{i+1}^*(x_i^*, y_i^*)$. The global error (GMAE) is computed as follows:

$$GMAE = D_G + MAE \tag{11}$$

Parameter Setting: In this paper, two kinds of regressions are used, coarse regression and fine regression. For coarse regression, 20 randomly selected faces are used as initialization and cascade level T = 18; for fine regression, 20 similar faces are used as initialization and cascade level T = 20. Each level contains K = 500 weak regressors and the depth of the tree used in regressor is D = 5. Shrinkage factor is 0.05. Following the feature selection constrain, we use 400 pairwise pixels and threshold corresponding to each pair is randomly chosen. For node splitting, we repeat S = 500 times to find the best one. The feedback procedure is repeated N_{re} = 1 time.

4.1 Comparison with Other Work

Table 2 Results of averaged error (%) 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 published papers.

Method	Common	Challenging	Fullset
DRMF [43]	6.65	19.79	9.22
ESR [9]	5.28	17.00	7.58
RCPR [41]	6.18	17.26	8.35
SDM [10]	5.57	15.40	7.50
ERT [11]	-	-	6.40
LBF [25]	4.95	11.98	6.32
cGPRT [27]	4.46	10.85	5.71
CFSS [28]	4.73	9.98	5.76
TCDCN [34]	4.80	8.60	5.54
RAR [44]	**4.12**	**8.35**	**4.94**
RDR [45]	5.03	8.95	5.80
Our method	4.26	8.50	5.09

Table 2 displays comparisons with state-of-the-art methods on 300W dataset. Compared methods include DRMF [43], ESR [9], RCPR [41], SDM [10], ERT [11], LBF [25], cGPRT [27], CFSS [28], RAR [44], RDR [45] and TCDCN [34]. From this Table, we can give the following conclusions: ESR [9], ERT [11] and cGPRT [27] are approaches of classical cascade shape regression framework, our method has a large improvement over these methods; RAR [44], RDR [45] and TCDCN [34] are state-of-the-art deep learning based methods, our results are also comparable with deep learning based methods. 300W challenging subset is a challenging dataset that has many faces with large variations on pose and expression. Our method makes a significant improvement on this dataset compares with the other traditional methods.

Table 3 shows the results of GMAE, structure distance D_G and computing time between different methods. The GMAE is an evaluation that measures the global distance between two faces. In this new metric, our method is much better than others. We use the speed that is provided in the corresponding paper. RAR [44] method takes 250 ms to compute a 256×256 face image and its speed is 4 fps. Our method is real time and much faster than RAR [44].

Table 3 Results of GMAE, D_G and computing speed (fps) compared with state-of-the-art approaches on 300W Dataset.

Method	GMAE	D_G	Speed
ESR [9]	11.51	3.75	350
ERT [11]	9.40	2.98	**1000**
LBF [25]	9.33	2.93	320
CFSS Practical [28]	8.84	2.91	24
cGPRT [27]	8.40	2.69	93
Our method	**7.23**	**2.14**	50

Fig. 5 and Fig. 6 show the comparison of CED and global CED curves with state-of-the-art approaches on 300W fullset, the following approaches includes DRMF [43], ESR [9], LBF [25], CFSS Practical [28] and cGPRT [27]. We can see that our approach gaps others in both CED and global CED curves. ESR [9] is reproduced by ourselves with the error of 7.76. The result of LBF [25] is provided by the author, the codes of DRMF [43] and CFSS Practical [28] are downloaded online.

Fig. 7 gives the comparison with 6 methods on 300W challenge subset and our approach has a significant improvement. We also show some visible results of 300W datasets in Fig. 9. Though it is difficult to detect the landmarks in the images, our method achieves good performance by applying the proposed method. Images in Fig. 8 are some frontal faces with different color shin, expression and illumination. The results show that our method is robust on these situations too.

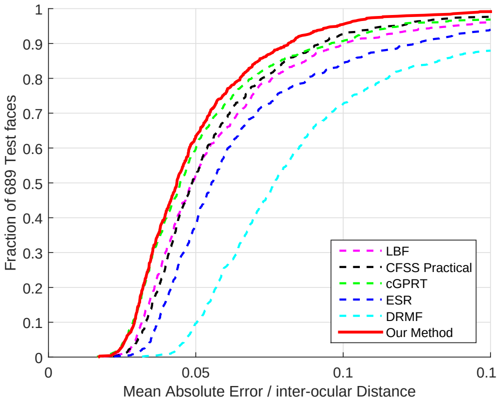

Fig. 5: Comparison of CED curves on 300W dataset.

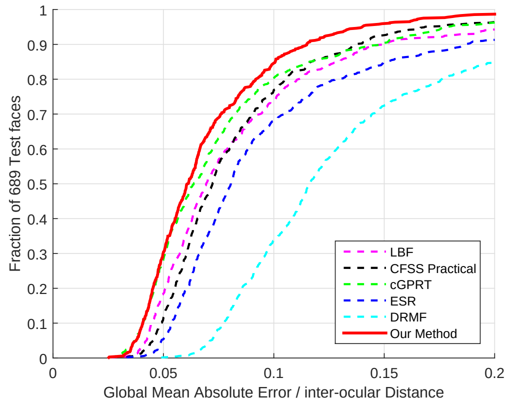

Fig. 6: Comparison of global CED curves on 300W dataset.

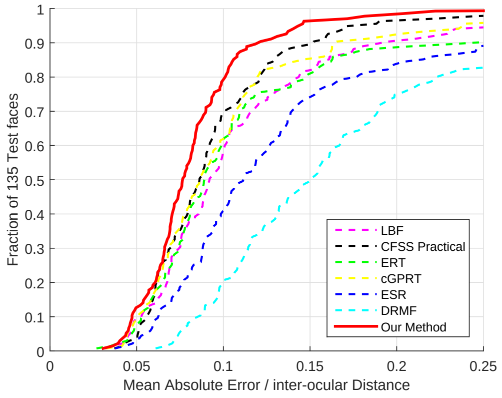

Fig. 7: Comparison of CED curves on 300W challenging dataset.

Table 4 Results of averaged error (%) compared with state-of-the-art approaches on COFW dataset.

Method	COFW
ESR [9]	11.2
RCPR [41]	8.50
SDM [10]	9.33
TCDCN [34]	8.05
RPP [46]	7.52
RAR [44]	6.03
PCD-CNN [47]	5.77
Our method	**5.23**

Comparison with state-of-the-art approaches on COFW dataset is showed in Table 4. COFW dataset is very challenging due to lots of faces with heavy occlusions. We report the results of some methods including ESR [9], RCPR [41], SDM [10], TCDCN [34], RPP [46], RAR [44] and PCD-CNN [47]. From this Table, we can see that our method is much better than other methods. With the help of the salient-to-inner-to-all manner, our method is robust on the conditions of occlusion. Fig. 10 shows some results of COFW dataset and demonstrates the availability of our method. In this figure, a large part of faces are invisible, our method can give good results with the help of the proposed method, especially salient landmark detection, the effect of occlusion is suppressed.

The experiments on 300W benchmark dataset show that our method is second best and closes to the best method. On COFW

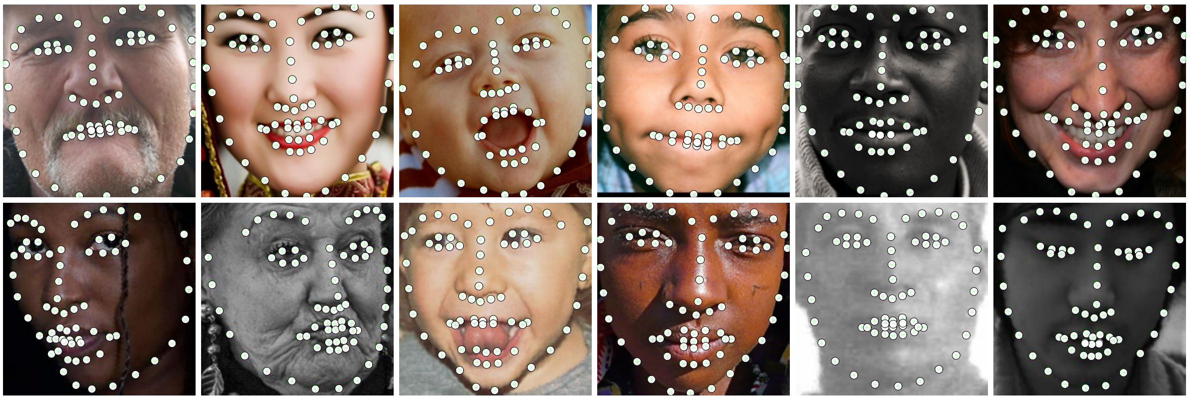

Fig. 8: Results of some frontal faces on 300W dataset.

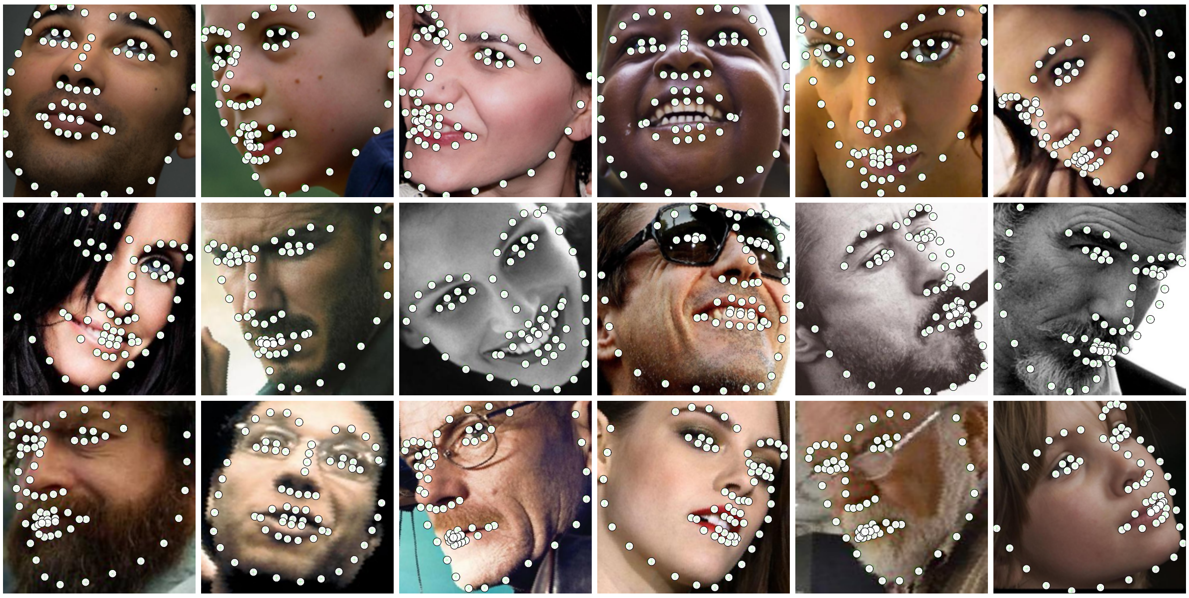

Fig. 9: Some challenging results of our method on 300W dataset.

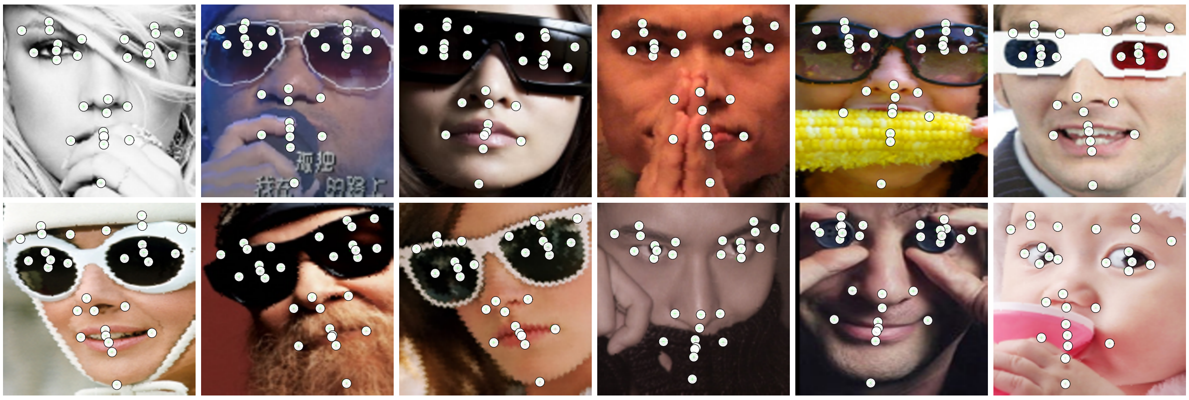

Fig. 10: Experimental results of our method on COFW dataset.

benchmark dataset, our method is much better than others. Deep learning methods are high accurate, but these methods need huge computing resource, millions of training data and long running time. Our method is based on linear regression framework which is efficient and need low computing resource. RAR [44] method takes 250ms to process an image of 256×256, our method takes 20ms. The memory footprint of our method is 300 MB and the memory footprint of RAR [44] is over 1GB. From the above, we can see that our method is comparable with state-of-the-art deep learning-based methods.

4.2 *Further Analyses*

This paper mainly copes with two problems of cascade shape regression, the unified initialization and regression structure. In the section of the introduction, we know that if an initial shape is similar to the input face, the algorithm can give better result. For instance, if we use an averaged face as initial face (like LBF [25]), the error of the common set is much lower than challenging set. That is because most face shapes in the common set are very similar to the average face. We utilize the 68 landmarks initial shapes as the final results to evaluate the accuracy of initialization. If mean shape is used as the final result on 300W full set, the error is bigger than 20. In manner of salient-to-inner-to-all, the error of using generated discriminative initial face as the final result is 7.90. Combining this manner with pose-invariant searching, the error of the initialization is 7.16. By applying the whole procedure of our framework, the error of 68 initial landmarks is 6.48. Obviously, the initialization has a vast improvement and each component has a contribution. And we can see that salient landmarks detection has a potential to be more accurate because it is the coarse result. The feedback operation is used to solve this problem and Fig. 3 shows its effectiveness, especially when the salient landmarks detection is not accurate.

4.3 Computation Complexity

The computation complexity of our approach mainly contains two parts, searching the similar faces and regression procedure. The complexity of searching the similar faces is O(M_{all}), where M_{all} is the number of training samples. The main cost of this algorithm is the second part. The complexity of one regression procedure is O($TKDP$), where P is the number of landmarks. Our experiments are conducted on a single core Intel(R) Xeon(R) CPU E5-2630 v3 @2.4 GHz. In 300W (68 landmarks) full set, our approach achieves about 50 fps (frame-per-second).

5 Conclusion

In this paper, we propose a feedback cascade shape regression method that follows a salient-to-inner-to-all manner. Firstly, Salient landmarks are detected by coarse regression, then inner landmarks and all landmarks are detected by fine regression. Both inner landmarks detection and pose-invariant retrieval can help to search high-quality similar faces to generate the initial face for fine regression. The salient landmarks are updated by feedback operation and final results are obtained in the salient-to-inner-to-all manner. Pipeline of salient-to-inner-to-all is insensitive to heavy occlusion and pose-invariant retrieval is robust on large pose. After feedback operation, the algorithm restarts from a new initialization and the results are more close to target location. Experimental results demonstrate our approach is robust and accurate.

Acknowledgment

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).

6 References

1 Chen, C., Dantcheva, A., Ross, A. 'Automatic facial makeup detection with application in face recognition'. In: International Conference on Biometrics. (Madrid, Spain), June 2013. pp. 1–8

2 Qian, J., Yang, J., Xu, Y.: 'Local structure-based image decomposition for feature extraction with applications to face recognition', *IEEE Transactions on Image Processing*, 2013, **22**, (9), pp. 3591

3 Xiong, X., De, F. la Torre. 'Global supervised descent method'. In: Conference on Computer Vision and Pattern Recognition. (Boston, MA, USA), June 2015. pp. 2664–2673

4 Guo, D., Sim, T. 'Digital face makeup by example'. In: Computer Society Conference on Computer Vision and Pattern Recognition. (Miami, Florida, USA), June 2009. pp. 73–79

5 Liu, L., Xing, J., Liu, S., Xu, H., Zhou, X., Yan, S.: 'Wow! you are so beautiful today!', *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2014, **11**, (1s), pp. 20

6 Wu, T., Turaga, P., Chellappa, R.: 'Age estimation and face verification across aging using landmarks', *IEEE Transactions on Information Forensics and Security*, 2012, **7**, (6), pp. 1780–1788

7 Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T.: 'The painful face: pain expression recognition using active appearance models', *Image and Vision Computing*, 2007, **27**, (12), pp. 1788–1796

8 Ramirez.Rivera, A., Castillo, R., Chae, O.: 'Local directional number pattern for face analysis: Face and expression recognition', *IEEE Transactions on Image Processing*, 2013, **22**, (5), pp. 1740–1752

9 Cao, X., Wei, Y., Wen, F., Sun, J.: 'Face alignment by explicit shape regression', *International Journal of Computer Vision*, 2012, **107**, (2), pp. 117–190

10 Xiong, X., De la Torre, F. 'Supervised descent method and its applications to face alignment'. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. (Portland, OR, USA), june 2013. pp. 532–539

11 Kazemi, V., Sullivan, J. 'One millisecond face alignment with an ensemble of regression trees'. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (Columbus, OH, USA), june 2014. pp. 1867–1874

12 Zhu, H., Sheng, B., Shao, Z., Hao, Y., Hou, X., Ma, L.: 'Better initialization for regression-based face alignment', *Computers & Graphics*, 2018, **70**, pp. 261–269

13 Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: 'Active shape models-their training and application', *Computer vision and image understanding*, 1995, **61**, (1), pp. 38–59

14 Cristinacce, D., Cootes, T.F. 'Feature detection and tracking with constrained local models'. In: Proceedings of the British Machine Vision Conference. (Edinburgh, UK), September 2006. pp. 929–938

15 Wang, Y., Lucey, S., Cohn, J.F. 'Enforcing convexity for improved alignment with constrained local models'. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (Anchorage, Alaska, USA), june 2008. pp. 1–8

16 Lucey, S., Wang, Y., Cox, M., Sridharan, S., Cohn, J.F.: 'Efficient constrained local model fitting for non-rigid face alignment', *Image and Vision Computing*, 2009, **27**, (12), pp. 1804–1813

17 Saragih, J.M., Lucey, S., Cohn, J.F.: 'Deformable model fitting by regularized landmark mean-shift', *International Journal of Computer Vision*, 2011, **91**, (2), pp. 200–215

18 Zhu, X., Ramanan, D. 'Face detection, pose estimation, and landmark localization in the wild'. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. (Anchorage, Alaska, USA), June 2012. pp. 2879–2886

19 Mostafa, E., Ali, A.A., Shalaby, A., Farag, A. 'A facial features detector integrating holistic facial information and part-based model'. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. (Boston, MA, USA), June 2015. pp. 93–99

20 Cootes, T.F., Edwards, G.J., Taylor, C.J.: 'Active appearance models', *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, pp. 681–685

21 Matthews, I., Baker, S.: 'Active appearance models revisited', *International Journal of Computer Vision*, 2004, **60**, (2), pp. 135–164

22 Liang, L., Xiao, R., Wen, F., Sun, J. 'Face alignment via component-based discriminative search'. In: Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Proceedings, PartII. (Marseille, France), October 2008. pp. 72–85

23 Dollar, P., Welinder, P., Perona, P. 'Cascaded pose regression'. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition. (San Francisco, CA, USA), June 2010. pp. 1078–1085

24 Dantone, M., Gall, J., Fanelli, G., Van.Gool, L. 'Real-time facial feature detection using conditional regression forests'. In: Conference on Computer Vision and Pattern Recognition. (Providence, RI, USA), June 2012. pp. 2578–2585

25 Ren, S., Cao, X., Wei, Y., Sun, J. 'Face alignment at 3000 fps via regressing local binary features'. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. (Columbus, OH, USA), june 2014. pp. 1685–1692

26 Tzimiropoulos, G. 'Project-out cascaded regression with an application to face alignment'. In: IEEE Conference on Computer Vision and Pattern Recognition. (Boston, MA, USA), June 2015. pp. 3659–3667

27 Lee, D., Park, H., Yoo, C.D. 'Face alignment using cascade gaussian process regression trees'. In: IEEE Conference on Computer Vision and Pattern Recognition. (Boston, MA, USA), June 2015. pp. 4204–4212

28 Zhu, S., Li, C., Loy, C.C., Tang, X. 'Face alignment by coarse-to-fine shape searching'. In: IEEE Conference on Computer Vision and Pattern Recognition. (Boston, MA, USA), June 2015. pp. 4998–5006

29 Zhang, J., Shan, S., Kan, M., Chen, X. 'Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment'. In: Computer Vision - ECCV 2014 - 13th European Conference, Proceedings, Part II. (Zurich, Switzerland), September 2014. pp. 1–16

30 Xiao, S., Yan, S., Kassim, A.A. 'Facial landmark detection via progressive initialization'. In: IEEE International Conference on Computer Vision Workshop, ICCV Workshops. (Santiago, Chile), December 2015. pp. 986–993

31 Sun, Y., Wang, X., Tang, X. 'Deep convolutional network cascade for facial point detection'. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. (Portland, OR, USA), June 2013. pp. 3476–3483

32 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. (Sydney, Australia), Decmember 2013. pp. 386–391

33 Honari, S., Yosinski, J., Vincent, P., Pal, C. 'Recombinator networks: Learning coarse-to-fine feature aggregation'. In: Conference on Computer Vision and Pattern Recognition. (Las Vegas, NV, USA), June 2016. pp. 5743–5752

34 Zhang, Z., Luo, P., Loy, C.C., Tang, X.: 'Learning deep representation for face alignment with auxiliary attributes', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**, (5), pp. 918–930

35 Friedman, J.H.: 'Greedy function approximation: a gradient boosting machine', *Annals of statistics*, 2001, pp. 1189–1232

36 Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: 'Fast keypoint recognition using random ferns', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**, (3), pp. 448–461

37 Liao, S., Jain, A., Li, S.: 'A fast and accurate unconstrained face detector', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **38**, (2), pp. 211–223

38 Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: 'Localizing parts of faces using a consensus of exemplars', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**, (12), pp. 545 – 552

39 Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T. 'Interactive facial feature localization'. In: 12th European Conference on Computer Vision, Florence, Proceedings, Part III. (Florence, Italy), October 2012. pp. 679–692

40 Messer, K., Matas, J., Kittler, J., Jonsson, K. 'Xm2vtsdb: The extended m2vts database'. In: Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA). (Washington D. C. USA), March 2000. pp. 72–77

41 Burgos-Artizzu, X.P., Perona, P., , Dollár, P. 'Robust face landmark estimation under occlusion'. In: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013. (Sydney, Australia), December 2013. pp. 1513–1520

42 Hao, Y., Zhu, H., Wu, K., Lin, X., Ma, L.: 'Salient-points-guided face alignment', *Multimedia Systems*, 2017, , (12), pp. 1–11

43 Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M. 'Robust discriminative response map fitting with constrained local models'. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. (Portland, OR, USA), June 2013. pp. 3444–3451

44 Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A. 'Robust facial landmark detection via recurrent attentive-refinement networks'. In: Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part I. (Amsterdam, The Netherlands), October 2016. pp. 57–72

45 Xiao, S., Feng, J., Liu, L., Nie, X., Wang, W., Yan, S., et al. 'Recurrent 3d-2d dual learning for large-pose facial landmark detection'. In: IEEE International Conference on Computer Vision. (Venice, Italy), October 2017. pp. 1642–1651

46 Yang, H., He, X., Jia, X., Patras, I.: 'Robust face alignment under occlusion via regional predictive power estimation.', *IEEE Transactions on Image Processing*, 2015, **24**, (8), pp. 2393–403

47 Kumar, A., Chellappa, R. 'Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment'. In: IEEE Conference on Computer Vision and Pattern Recognition. (Salt Lake City, Utah, USA), June 2018.