Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment

Jie Zhang[1,2], Shiguang Shan[1], Meina Kan[1], and Xilin Chen[1]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
{jie.zhang,shiguang.shan,meina.kan,xilin.chen}@vipl.ict.ac.cn

Abstract. Accurate face alignment is a vital prerequisite step for most face perception tasks such as face recognition, facial expression analysis and non-realistic face re-rendering. It can be formulated as the nonlinear inference of the facial landmarks from the detected face region. Deep network seems a good choice to model the nonlinearity, but it is non-trivial to apply it directly. In this paper, instead of a straightforward application of deep network, we propose a Coarse-to-Fine Auto-encoder Networks (CFAN) approach, which cascades a few successive Stacked Auto-encoder Networks (SANs). Specifically, the first SAN predicts the landmarks quickly but accurately enough as a preliminary, by taking as input a low-resolution version of the detected face holistically. The following SANs then progressively refine the landmark by taking as input the local features extracted around the current landmarks (output of the previous SAN) with higher and higher resolution. Extensive experiments conducted on three challenging datasets demonstrate that our CFAN outperforms the state-of-the-art methods and performs in real-time(40+fps excluding face detection on a desktop).

Keywords: Face Alignment, Nonlinear, Deep Learning, Stacked Auto-encoder, Coarse-to-Fine, Real-time.

1 Introduction

Face alignment or facial landmark detection plays an important role in face recognition, facial expression recognition, face animation, *etc.* Therefore, it has received more and more attentions in recent years. However, it remains a challenging problem due to the complex variations in face appearance caused by pose, expression, illumination, partial occlusion, *etc.* Generally speaking, the existing approaches can be categorized into holistic feature based methods [7,21,14,34,19,6] and local feature based methods [8,10,15,23,9,25,35,32,31,2,28,11].

As a typical model, Active Appearance Models (AAM) [7,21] firstly use Principal Component Analysis (PCA) to model the shape and texture separately and then integrate them together with another PCA to get the generative appearance model. In the testing stage, the shape of a new face image is inferred by optimizing the model parameters to minimize the difference between the observed

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 1–16, 2014.

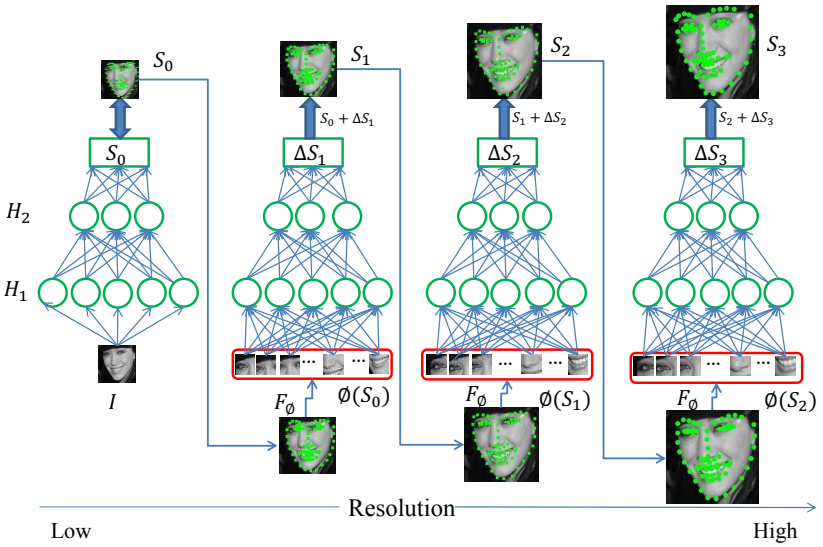


Fig. 1. Overview of our Coarse-to-Fine Auto-encoder Networks (CFAN) for real-time face alignment. H_1, H_2 are hidden layers. Through function F_Φ, the joint local features $\Phi(S_i)$ are extracted around facial landmarks of current shape S_i.

face image and the image generated by the appearance model. However, these methods generally fail in case of complex appearance variations in real-world applications, mainly because a single linear model can hardly cover all the nonlinear variations in facial appearance. To address this problem, Zhao et al. [34] propose a locally linear AAM method to approximate the global nonlinear model and report good performance. However the initialization of this approach needs eyes locations. Moreover, it is difficult for AAM-like holistic methods to handle partial occlusion problems.

Instead of modeling appearance with the entire face, local feature based methods like ASMs [8,15,23], CLM [9] build appearance models with local image patches which are generally sampled around the current facial landmarks. In these methods, partial occlusion problem can be easily handled by including a shape constraint. But the shape constrains employed in these methods are relatively weak so that they are prone to local minimum due to ambiguous local regions [8,9]. Saragih et al. [25] propose a Regularized Landmark MeanShift fitting method to solve the optimization problem of CLM, which achieves higher performance for generic face alignment scenario. Recently, Asthana et al. propose a promising method named as discriminative response map fitting with constrained local models (DRMF) [2], which learns the dictionaries of probability response maps based on local features and adopts linear regression-based fitting method in the CLM framework. In another state-of-the-art work [31], a

Fig. 2. Facial landmark detection under the partial occlusion scenario (from Helen datasets [18]): Results of DCNN [26] (top row) and our CFAN (bottom row)

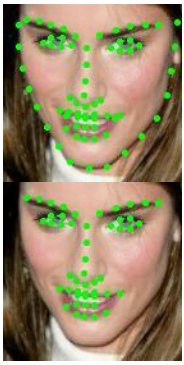


Fig. 3. Denition of 68 (top) and 49 (bottom) facial landmarks

supervised descent method (SDM) is proposed to solve nonlinear least squares optimization problem and achieves significant success in facial landmark detection.

SDM [31] achieves promising performance by using the supervised descent strategy, but it is initialized by using the mean shape on the detected face region which makes it heavily rely on the face detection results. Moreover, in each stage of the cascades architecture of SDM, linear regression is exploited to model the mapping from shape-index feature [12] to face shape, which may be insufficient for the complex non-linear process from shape-index feature to face shape. In contrast, the DCNN [26] employs a deep CNN model with the global feature to predict the landmark locations as the initialization, which is more accurate than the mean shape. After the initialization, the successive network of DCNN refines each landmark separately without any shape constraints, which may fail in case of partial occlusions as shown in the first row of Fig. 2. Therefore it is necessary for the first stage to provide a strong shape prior for the following stages.

In this paper, we further push the frontier of the area by resorting to deep network and elaborately adapting it to disintegrate progressively the complex nonlinearity in face shape inference. We propose an architecture named Coarse-to-Fine Auto-encoder Networks (CFAN), as illustrated in Fig. 1, and show how it can further beat the state-of-the-art methods such as SDM and DRMF. As seen from Fig.1, instead of a single stacked auto-encoder network (SAN), our CFAN is comprised of several successive SANs, each figuring out part of the nonlinearity. Specifically, the first SAN predicts the face shape quickly by taking holistically a low-resolution version of the detected face as input; the following SANs then progressively refine the landmark locations by taking as input the joint local features extracted around the current landmarks (output of the previous SAN) in higher and higher resolution. By using such a progressive and resolution-variable

strategy, the search space of each SAN, or in other words the difficulty of the task for each SAN, is well controlled and thus more tractable. Benefitted from the advantages of joint local features, our method is more robust to partial occlusions than DCNN [26] as shown in the last row of Fig. 2.

Extensive evaluation results on several public databases, *i.e.*, XM2VTS [22], LFPW [3] and HELEN [18], show that our method achieves impressively better accuracies, compared with the state-of-the-art methods, such as SDM and DRMF. Furthermore, our method (in Matlab codes) takes about 23 milliseconds per image to predict 68 facial points excluding the face detection time, on an desktop machine with Intel i7-3770 (3.4 GHz CPU).

2 Related Works

2.1 Local Models with Regression Fitting

Recently local model methods with Regression Fitting [28,11,33,31,2] make great progresses on facial point detection, especially SDM [31]. Local methods like ASMs [8,15,23] and CLMs [9,25] solve the optimization problems with Gauss-Newton method. Yet, instead of computing the Jacobian and Hessian matrices, SDM learns generic descent directions and re-scaling factors by using the linear regression. Specifically, given an image $x \in \mathbf{R^d}$, S denotes the shape vector containing the coordinates of the facial points. The objective of most regression fitting model can be formulated as optimizing a sequence of successive update ΔS for shape as follows:

$$f(S_0 + \Delta S) = ||\Phi(S_0 + \Delta S) - \Phi(S_g)||_2^2, \tag{1}$$

where S_0 and S_g denote the initial shape and ground truth shape respectively and Φ is a nonlinear feature extraction function from a shape. The shape update ΔS can be obtained by employing Newton's method as follows:

$$\Delta S = -H^{-1} J_f = -2H^{-1} J_\phi^{\mathrm{T}}(\Phi(S_0) - \Phi(S_g)), \tag{2}$$

where J_f and H are the Jacobian and Hessian matrices.

SDM directly estimates the descent direction $R_1 = -2H^{-1}J_\phi^{\mathrm{T}}$ by using a linear regression between the appearance information and the shape deviation to avoid the complex computations of Jacobian and inverse of Hessian matrices. Thus, in SDM, Eq. (2) is formulated as bellow:

$$\Delta S_1 = R_1 \Phi_0 + b_1, \tag{3}$$

where b_1 is a bias term corresponding to $\Phi(S_g)$. In a similar way, SDM can learn a sequence of generic descent directions R_k and bias term b_k after k iterations.

$$\Delta S_k = R_k \Phi_{k-1} + b_k. \tag{4}$$

For most methods including SDM, the mean shape is used as the initialization, which may suffer from local minimum problem in case of bad initializations. To depress the effects from bad initializations, Cao et al. [6] use multiple initializations strategy and Burgos-Artizzu et al. [5] adopt smart restarts technique, but it still leaves a long way to go.

2.2 Deep Models

Recently, deep models like Deep Auto-encoders(DAEs), Convolutional Neural Networks(CNNs), Restricted Boltzmann Machines(RBMs) and their variants are widely used in the field of computer vision [4]. They have achieved great success in many challenging tasks such as image classification [17], scene parsing [13], human pose estimation [27], face alignment and facial feature tracking [26,30].

Sun et al. [26] propose a cascaded regression approach for facial points detection with three-stage deep convolutional network. At the first stage, the carefully designed convolutional neural networks provide accurate initial estimations of facial points when given the full face as input. Then the initial estimations are refined during next two stages. Impressive results are achieved on two public datasets, BioID [16] and LFPW [3]. However, in the layers after the first one, each landmark is refined separately, which makes it depend on and sensitive to the accuracy of the first layer more heavily. Another interesting work [30] constructs a face shape prior model by using RBMs and their variants for facial feature tracking under varying facial expressions and face poses. In [30], Wu et al. use deep belief networks(DBNs) to capture the face shape variations from facial expressions and handle pose variations with a 3-way RBM model. Luo et al. [20] also use DBNs for facial component detection and then train the facial component segmentators with deep auto-encoders.

Most of these deep models achieve promising results on facial landmarks detection and tracking, benefitted from its favorable ability for modeling the nonlinearity, which can work well for the nonlinear mapping from the a face image to the face shape. Some major concerns in these deep works are the time complexity and the local minima, due to the highly nonlinear optimization.

3 Coarse-to-Fine Auto-Encoder Networks

In this paper, we present a novel Coarse-to-Fine Auto-encoder Networks method (CFAN) for real-time facial landmark detection. Firstly, we will illustrate the overview of the proposed framework; secondly, we will describe the details about two components of CFAN, *i.e.*, global SAN and local SANs; and finally we will give a detailed discussion about the difference from some existing works.

3.1 Method Overview

As shown in Fig. 1, the proposed CFAN attempts to design the general cascade-regression framework in a coarse-to-fine architecture, with the regression in each stage modeled as a nonlinear deep network. Specifically, the CFAN framework consists of several successive Stacked Auto-encoder Networks (SANs). Each SAN attempts to characterize the nonlinear mappings from face image to face shape in different scales based on the shape predicted from the previous SAN.

The first SAN (referred as global SAN) endeavors to roughly approximate the facial landmark locations, and therefore a low-resolution image is exploited for

a large search step. A large step can alleviate the suffering from local minima and meanwhile promise a fast model. Moreover, rather than local shape-indexed feature from mean shape, the global image feature is employed as input to avoid the inaccuracy of mean shape. As a result, the global SAN can approach the ground truth facial landmark locations more accurately and more quickly.

After getting an estimation S_0 of face shape from the first SAN, the successive SANs (referred as local SANs) make an effort to refine the shape by regressing the deviation ΔS between the current locations and the ground truth locations step by step. The nonlinear regression model SAN is still exploited to model the nonlinearity between the current feature and the ground truth shape. To characterize fine variations, the shape-indexed feature extracted from current shape at higher resolution is exploited to enforce smaller search step and smaller search region. Furthermore, the shape-indexed features of all facial points are concatenated together to enforce all facial points updated jointly so as to insure a reasonable solution, even under the partial occlusion scenario.

3.2 Global SAN

The first SAN of the proposed coarse-to-fine deep networks, *i.e.*, the global SAN, directly estimates the face shape based on global raw features at a low-resolution image. Given a face image $x \in \mathbf{R^d}$ of d pixels, $S_g(x) \in \mathbf{R^p}$ denotes the ground truth locations of p landmarks. The face landmark detection is to learn a mapping function $\mathbf{F}$ from the image to the face shape as follows:

$$\mathbf{F} : S \leftarrow x. \tag{5}$$

Generally, $\mathbf{F}$ is complex and nonlinear. To achieve this goal, k single hidden layer auto-encoders are stacked as a deep neural network to map the image to the corresponding shape. Specifically, the face alignment task is formulated as minimizing the following objective:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} ||S_g(x) - f_k(f_{k-1}(...f_1(x)))||_2^2, \tag{6}$$

$$f_i(a_{i-1}) = \sigma(W_i a_{i-1} + b_i) \triangleq a_i, i = 1, ..., k-1, \tag{7}$$

$$f_k(a_{k-1}) = W_k a_{k-1} + b_k \triangleq S_0. \tag{8}$$

where $\mathbf{F} = \{f_1, f_2, ..., f_k\}$, f_i is the mapping function of i^{th} layer in the deep network, σ is a sigmoid function and a_i is the feature representations of each layer. Nonlinear mapping in term of sigmoid function is employed by the first $k-1$ layers to characterize the nonlinearity between the image feature and the face shape. However, the output range of sigmoid function is [0 1] which is inconsistent with the location range, therefore, linear regression is exploited in the last layer f_k to get an accurate shape estimation S_0.

To prevent over-fitting, a regularization term $\sum_{i=1}^{k} ||W_i||_F^2$(a weight decay term) is added which tends to decrease the magnitude of the weights. The objective function is further re-formulated as bellow:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} ||S_g(x) - f_k(f_{k-1}(...f_1(x)))||_2^2 + \alpha \sum_{i=1}^{k} ||W_i||_F^2. \quad (9)$$

The function $\mathbf{F}$ contains lots of parameters and it is easy to fall into local minimum during optimization. To achieve a better optimization, firstly, we adopt the unsupervised pre-train process to initialize the first $k-1$ layers in a stacked strategy and random initialization for the k^{th} layer; secondly, fine tune the whole network in a supervised way.

For the i^{th} layer, it is pre-trained by optimizing the following objective function:

$$\{f_i{}^*, g_i{}^*\} = \arg\min_{f_i, g_i} ||a_{i-1} - g_i(f_i(a_{i-1}))||^2 + \alpha(||W_i||_F^2 + ||W_i^T||_F^2), \quad (10)$$

where $f_i(x) = \sigma(W_i x + b_i)$, $g_i(x) = \sigma(W_i^T x + b_i')$, $i = 1, 2, ..., k-1$.

Then the output of this single hidden layer network $a_i = f_i(a_{i-1})$ is used as the input of the next layer. For the first layer, the input is the raw image feature, *i.e.*, $a_0 = x$.

After the initialization with Eq. (10), all layers of the whole network are fine-tuned according to Eq. (9). As a result, the first few layers of a stacked auto-encoder network tend to capture the low-level features such as texture patterns in an image, while the higher layers tend to capture higher-level features containing context information of texture patterns.

After the optimization, the prediction of the facial landmarks is achieved as S_0, which is a rough but robust and fast approximation of the ground truth.

3.3 Local SANs

The global SAN described above will give a rough shape estimation S_0 of input image x, which is already close to the ground truth locations but not close enough due to the highly complicated variations in expression, pose, identity, *etc.* To achieve finer locations, several successive SANs are employed to iteratively predict the deviation ΔS_j between current shape S_{j-1} and the ground truth S_g based on joint local shape-indexed features, referred as local SANs.

Shape-indexed features extracted around the landmark points have been proved to be efficient and effective for face alignment [12,6,31,5]. The local feature from each facial point can only capture the information from itself while ignore the relevance with the other points. Therefore, the facial points are modeled jointly in our local SAN, by concatenating all local shape-indexed feature together as the input.

Similarly as the global SAN, the successive local SAN is also designed as a stacked deep auto-network to deal with the nonlinearity of predicting the face shape, but with the local shape-indexed feature as input. With the estimated

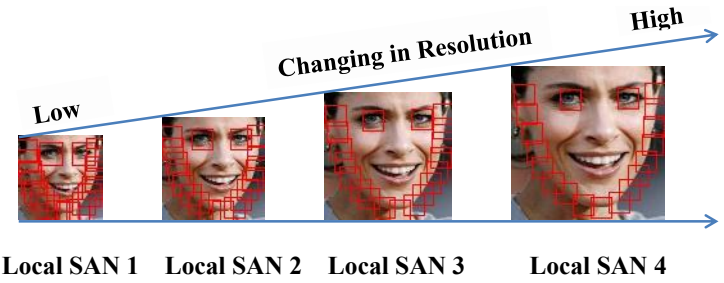


Fig. 4. Local patches extracted around the landmark points with different resolutions. For the sake of concise display, we choose two eye centers and 17 facial points on the face contour to describe the multi-resolution strategy used in each local SAN.

shape S_0 from global SAN, the shape-indexed features, *i.e.*, SIFT, can be extracted around each facial point, denoted as $\phi(S_0)$. The objective of the first local SAN is to achieve a nonlinear regression $\mathbf{H}_1$ from the shape-indexed feature $\phi(S_0)$ to the deviation $\Delta S_1 = S_g - S_0$ as follows:

$$\mathbf{H}_1^* = \arg\min_{\mathbf{H}_1} ||\Delta S_1(x) - h_k^1(h_{k-1}^1(...h_1^1(\phi(S_0))))||_2^2 + \alpha \sum\nolimits_{i=1}^{k} ||W_i^1||_F^2, \quad (11)$$

where $\mathbf{H}_1 = \{h_1^1, h_2^1, ..., h_k^1\}$, $\sum_{i=1}^{k} ||W_i^1||_F^2$ is the weight decay term. Similar to the global SAN, the whole deep network is firstly initialized by using the unsupervised pre-training, and then fine-tuned according to Eq. (11).

After getting the face shape update ($\Delta \hat{S}_1$) by the first local SAN, an updated face shape can be obtained as $S_1 = S_0 + \Delta \hat{S}_1$. Then the successive local SAN extracts local features around the new shape, and optimizes a deep network to minimize the new deviation between the current location and the ground truth. The objective of the j^{th} local SAN is shown as follows:

$$\mathbf{H}_j^* = \arg\min_{\mathbf{H}_j} ||\Delta S_j(x) - h_k^j(h_{k-1}^j(...h_1^j(\phi(S_{j-1}))))||_2^2 + \alpha \sum\nolimits_{i=1}^{k} ||W_i^j||_F^2. \quad (12)$$

For each Local SAN, local features around the landmark points are extracted in a local patch of the same size but at different resolutions as shown in Fig. 4. Local patches of the same size at low-resolution face images contain more context information and thus lead to a larger searching region for the Local SAN. It is necessary for the anterior SANs to approximate with a large search step when the current location is relatively far from the ground truth. On the other hand, the local patches of the same size but at high resolution face images actually constrain the searching within a small region which means that the posterior SAN can refine the location with a tiny step leading to more accurate results.

3.4 Discussions

Differences with SDM [31]. A sequence of generic descent directions are learned by several successive local SANs as well as SDM [31], but they differ in the following aspects: 1) SDM employs linear regression to model the mapping from shape-indexed features to a face shape, while our CFAN employs nonlinear regression, *i.e.*, deep auto networks, to model the mapping from shape-indexed feature to face shape, which can achieve lower regression error. 2) SDM employs the mean shape as the initialization of the shape-indexed feature, which may be trapped when the initialization is far away from the ground truth, especially under the linear model. On the contrary, our CFAN designs a deep auto network to directly predict a rough estimation of the face shape from the global image feature rather than shape-indexed feature, and this can obtain a more accurate initialization of the shape for the following local SANs.

Differences with DCNN [26]. Both DCNN and our CFAN follow the cascade framework and use a global nonlinear regression as the first stage to achieve a rough estimation of face shape. The differences lie on two aspects: 1) In DCNN, after the global estimation, each facial point is refined independently, which may distort the whole shape without the constraint between facial points. On the contrary, all facial points are refined jointly in our CFAN and this can ensure an effective shape, especially when several landmarks are occluded in which case the rest will provide supports for locating the obscured one. 2) The separate refinement of each point makes DCCN framework rely on and sensitive to the accuracy of the first level more heavily than ours.

4 Implementation Details

Data Augmentation. To train a robust global SAN model, we augment the training data by perturbing each training sample with random changes in translation, rotation and scaling. This can effectively prevent the deep models from over-fitting and achieve robustness to various changes in the wild data.

Parameter Setting. The global SAN has four layers with three hidden layers followed by a linear regression layer that is capable of learning non-linear mappings from a full face with 50×50 pixels to a face shape. Numbers of hidden units in each layer are respectively $1600, 900, 400$. For local SANs, SIFT features are extracted around each landmark. The resolution of face images in each layer becomes higher and higher gradually during the successive local SANs. Numbers of hidden units in each layer of local SAN are respectively $1296, 784, 400$. The weight decay parameter α controls the relative importance of the two terms, the average sum-of-squares error term and the weight decay term. Although α can be set different, the same value $\alpha = 0.001$ is used for both global SAN and local SANs for simplicity.

5 Experiments

In this section, we firstly illustrate the experimental settings for the evaluations including the datasets and methods to compare; and then investigate the alignment results of each stage in our method; finally, compare the proposed CFAN with the state-of-the-art methods.

5.1 Datasets and Methods for Comparison

To evaluate the effectiveness of the proposed CFAN algorithm, four public datasets are used for our experiments, *i.e.*, XM2VTS [22], LFPW [3], HELEN [18] and AFW [35]. The images in XM2VTS dataset are collected under laboratory conditions, while the images in LFPW, HELEN and AFW datasets are collected in the wild environment formulating a more challenging scenario than XM2VTS. Face detection results can be achieved from ibug websit [1], and the ground truth annotations of 68 facial points (as shown in Fig. 3) are provided by [24].

We evaluate a few state-of-the-art methods, *i.e.*, DRMF [2], SDM [31], Zhu et al. [35] and Yu et al. [32]. For Zhu et al.'s method, we use the model released by Asthana et al. [2], which performs better as illustrated in [2]. The 68 facial landmarks predicted by Zhu et al. [35] are shown in Fig 3. Both of the publicly available codes from [2] and [32] predict 66 facial points (as shown in Fig. 3 except two inner mouth corners), and the released code of SDM only estimates 49 landmarks (as shown in Fig. 3) located in the inner regions of the face. For fair comparisons with these methods, we implement the SDM algorithm to estimate 68 points using the same training set, among which the common 66 facial points are used for evaluation. The normalized root mean squared error (NRMSE) is employed to measure the error between the estimated facial landmark locations and the ground truth. The NRMSE is normalized by the distance between centers of eyes. The cumulative distribution function (CDF) of NRMSE is applied for performance evaluation.

5.2 Investigation of Each SAN in CFAN

As the proposed CFAN method consists of several successive SANs, we investigate how each SAN contributes to the performance improvement for facial landmark detection. The experiments are conducted on LFPW dataset in terms of average detection accuracy of 66 facial points, *i.e.*, CDF. The images from LFPW training set [3], HELEN [18] and AFW [35] are used for training and the images in LFPW test set [3] are used for evaluation.

The evaluation results are shown in Fig. 5. As seen, the CDF of global SAN is 0.65 when NRMSE is 0.1, which is much better than mean shape. However, the estimated shape is still far away from the ground truth since global SAN just gives a roughly accurate estimation of facial landmark locations. But benefited from this more accurate shape estimation rather than the mean shape as shape initialization for local SANs, accuracy of facial landmark detection is significantly

improved by 25% in the first local SAN. In the second local SAN, the detection accuracy is improved up to about 5.7% when NRMSE is 0.1 and 44% when NRMSE is 0.05. In the third local SAN, no improvement when NRMSE is 0.1, and the improvement is up to 11% when NRMSE is 0.05. It demonstrates that the former SANs mainly handle the large variations due to pose and expression and the latter ones precisely refine the landmark locations in a smaller search region as the resolution becomes higher and higher.

Besides the performance, another important factor is the time complexity. We evaluate the run time of each SAN on LFPW as shown in Table 1. The method is run in matlab 2012 on a desktop (Intel i7-3770 3.4 GHz CPU). To avoid the influence of random factors, the method is repeated several times, and the average of running time is reported. As shown in the table, it takes only 0.25 millisecond per image for global SAN to give a rough estimation of the face shape. Each local SAN costs about 7 milliseconds and only 3 local SANs are enough. So, totally our CFAN takes less than 25 milliseconds per image for 68 facial points locating, which can easily meet the real-time requirement.

5.3 Comparison on XM2VTS Dataset

Firstly, we evaluate our CFAN and the existing methods under the controlled settings on XM2VTS dataset [22]. The XM2VTS dataset contains 2360 face images of 295 individuals collected over 4 sessions. In this experiment, our CFAN is trained by using the images from LFPW training set [3], HELEN [18] and AFW [35] and all methods are tested on XM2VTS. For DRMF method [2], the Viola-Jones face detector [29] is employed since all images in XM2VTS are almost frontal. For a fair comparison, only the common images with face detected by all methods are employed for the testing.

The cumulative error distribution curves of these methods are shown in Fig. 6. As seen, DRMF performs better than [35,32], followed by SDM which benefits from its supervised descent solution. Furthermore, our method performs the best on this dataset, even better than SDM, which attributes to the nonlinear model and coarse-to-fine strategy. The training set of our CFAN is composed of different datasets including large variations from pose, expression, illumination, partial occlusions *etc*, and while the major variation of XM2VTS is from the identity with similar pose, expression and illumination. This means that the distribution of training set of our CFAN is extremely different from the testing samples. Even trained from a different distribution, CFAN still works well, which demonstrate our method is robust to the out-of-database scenario.

Table 1. Run time of each stage in terms of millisecond (ms)

	Global SAN	Local SAN1	Local SAN2	Local SAN3	Total
Run Time (ms)	0.25	7.63	7.28	7.68	22.84

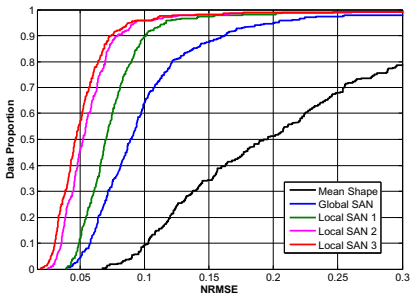


Fig. 5. The cumulative error distribution curves from LFPW of each SAN

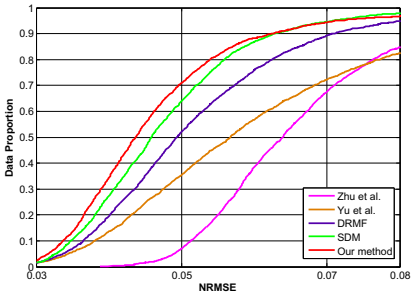


Fig. 6. Comparison on XM2VTS

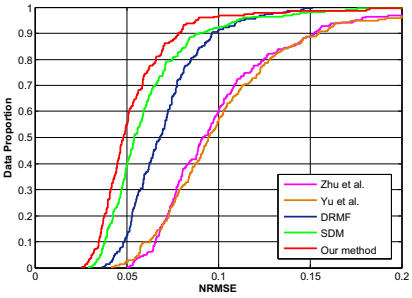


Fig. 7. Comparison on LFPW

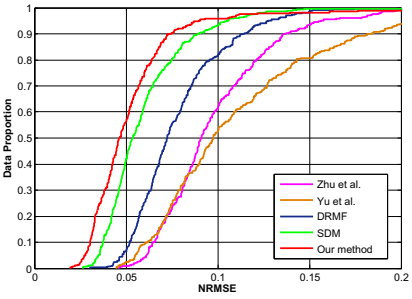
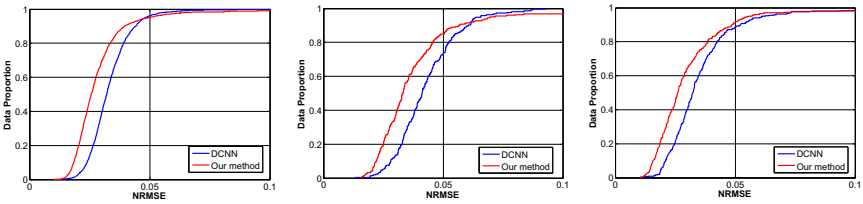


Fig. 8. Comparison on Helen

Furthermore, we compare our CFAN with DCNN [26] on this dataset in terms of five landmarks for a fair comparison since only the model of five landmarks is released. We directly run the model provided by [26] on XM2VTS dataset. The comparison results are shown in Fig. 9(a). As seen, our CFAN outperforms DCNN in general.

5.4 Comparison on LFPW Dataset

Furthermore, we evaluate the methods on the Labeled Face Parts in the Wild (LFPW) dataset [3] which is collected from wild condition. LFPW dataset consists of 1132 training images and 300 test images with large variations in pose, expression, illumination, partial occlusion, *etc*, which makes the facial point detection quite challenging on this dataset. The original URLs of images are provided by [3], but some of them are not available any longer. So, the 811 training samples and 224 test samples provided by ibug websit mentioned above are directly used for training and testing. For our method, we directly use the landmark detector trained for XM2VTS experiments for the evaluation. For DRMF method, the tree-based face detector is used to achieve more accurate face detection.



(a) XM2VTS Datasets (b) LFPW Datasets (c) Helen Datasets

Fig. 9. Comparison with DCNN [26]

The performance of all methods are shown in Fig. 7. As seen, SDM still performs the best among the existing methods, and our method achieves a better detection accuracy with an improvement up to about 15% than SDM when NRMSE is below 0.05. Similarly, we also compare our CFAN with DCNN [26] in terms of five landmarks. As seen from the Fig. 9(b), our CFAN performs much better than DCNN when NRMSE is below 0.06, but comparable or a little worse when NRMSE is 0.1. On average, our CFAN outperforms DCNN in terms of five landmarks, and a further significant improvement can be expected in terms of more points, *e.g.*, 68 points, especially considering those hard points around the contour.

5.5 Comparison on Helen Dataset

Similar to LFPW, the Helen dataset [18] is also collected under uncontrolled condition, *i.e.*, *Flicker*. Helen consists of 2330 high-resolution images with large variations in pose, lighting, expression, occlusion, and identity. For our CFAN, the images from the Helen training set, the LFPW training set [3], and AFW [35] are used for training the model. All methods are valuated on the 330 images from the Helen test set.

The comparison results are shown in Fig. 8. As seen, our CFAN still performs the best, which demonstrates the superiority to the existing methods again. As analyzed in Sec. 3.4, DCNN cannot well handle partial occlusion problem since each landmark is refined independently without any support from other points. Some failed examples are shown in Fig. 2. On the contrary, our method is more robust than DCNN under the partial occlusion scenario. Fig. 9(c) further shows that our CFAN performs better than DCNN on this dataset.

Fig. 10 shows the detection results of CFAN on some extremely challenging example faces from XM2VTS, LFPW and HELEN. It can be observed that our algorithm is robust to the variations from pose, expression, beard, sunglass and partial occlusion. However, as shown in the last column of Fig. 10, the performance of CFAN degrades on some images with simultaneous large out-of-plane rotations and exaggerated expressions, partially due to the lack of such samples in training set. Models specific to large pose or with latent pose estimation will be considered in the future.

Fig. 10. Example results from XM2VTS, LFPW and HELEN. The first five column samples contain diverse variations in pose, expression, beard, sunglass and occlusion respectively. Some failure cases are shown in the last column.

6 Conclusions and Future Works

Aiming at dealing with the nonlinearity in inferring face shapes from face images, we make use of a sequence of Stacked Auto-encoder Networks in a coarse-to-fine architecture, each of which figures out part of the nonlinearity. The first SAN takes directly a low-resolution version of the detected face as input, to globally estimate a roughly accurate shape. Then, the subsequent SANs take as input the shape-index local features at higher and higher resolution to refine the shape better and better. Such a coarse-to-fine strategy is proved well matching the capacity of SAN and the difficulty of the problem to solve, thus achieves better results than the state-of-the-art methods, such as SDM and DRMF, on three databases with extensive variations. Furthermore, our method can work rather efficiently, with 40+ fps even with Matlab codes on a common desktop with no parallel programming.

Our work further validates the effectiveness of regression-based methods for facial landmarks localization. By decomposing the nonlinearity of the image-to-shape mapping elaborately into a cascaded stages, facial landmarks can be accurately predicted progressively. In the future, we will try other types of deep networks with similar principle.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contracts Nos. 61390511, 61222211, 61173065, and 61272319.

References

1. 300 faces in-the-wild challenge, http://ibug.doc.ic.ac.uk/resources/300-W/
2. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3444–3451 (2013)
3. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 545–552 (2011)
4. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends® in Machine Learning 2(1), 1–127 (2009)
5. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: IEEE International Conference on Computer Vision, ICCV (2013)
6. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2887–2894 (2012)
7. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 23(6), 681–685 (2001)
8. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding (CVIU) 61(1), 38–59 (1995)
9. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: British Machine Vision Conference (BMVC), vol. 17, pp. 929–938 (2006)
10. Cristinacce, D., Cootes, T.F.: Boosted regression active shape models. In: British Machine Vision Conference (BMVC), pp. 1–10 (2007)
11. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2578–2585 (2012)
12. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085 (2010)
13. Grangier, D., Bottou, L., Collobert, R.: Deep convolutional networks for scene parsing. In: International Conference on Machine Learning Workshops, vol. 3 (2009)
14. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing (IVC) 23(12), 1080–1093 (2005)
15. Gu, L., Kanade, T.: A generative shape regularization model for robust face alignment. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 413–426. Springer, Heidelberg (2008)
16. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the hausdorff distance. In: International Conference on Audio-and Video-based Biometric Person Authentication (AVBPA), pp. 90–95 (2001)
17. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1106–1114 (2012)
18. 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, Part III. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012)

19. Liu, X.: Discriminative face alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 31(11), 1941–1954 (2009)
20. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2480–2487 (2012)
21. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision (IJCV) 60(2), 135–164 (2004)
22. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: The extended m2vts database. In: International Conference on Audio and Video-based Biometric Person Authentication (AVBPA), vol. 964, pp. 965–966 (1999)
23. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008)
24. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 896–903 (2013)
25. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: IEEE International Conference on Computer Vision (ICCV), pp. 1034–1041 (2009)
26. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3476–3483 (2013)
27. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2014)
28. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2729–2736 (2010)
29. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, p. I–511 (2001)
30. Wu, Y., Wang, Z., Ji, Q.: Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3452–3459 (2013)
31. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2013)
32. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: IEEE International Conference on Computer Vision, ICCV (2013)
33. Zhao, X., Kim, T.K., Luo, W.: Unified face analysis by iterative multi-output random forests. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2014)
34. Zhao, X., Shan, S., Chai, X., Chen, X.: Locality-constrained active appearance model. In: Asian Conference on Computer Vision (ACCV), pp. 636–647 (2013)
35. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886 (2012)