Occlusion-free Face Alignment: Deep Regression Networks Coupled with De-corrupt AutoEncoders

Jie Zhang[1,2] Meina Kan[1,3] Shiguang Shan[1,3] 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
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{jie.zhang,meina.kan,shiguang.shan,xilin.chen}@vipl.ict.ac.cn

Abstract

Face alignment or facial landmark detection plays an important role in many computer vision applications, e.g., face recognition, facial expression recognition, face animation, etc. However, the performance of face alignment system degenerates severely when occlusions occur. In this work, we propose a novel face alignment method, which cascades several Deep Regression networks coupled with De-corrupt Autoencoders (denoted as DRDA) to explicitly handle partial occlusion problem. Different from the previous works that can only detect occlusions and discard the occluded parts, our proposed de-corrupt autoencoder network can automatically recover the genuine appearance for the occluded parts and the recovered parts can be leveraged together with those non-occluded parts for more accurate alignment. By coupling de-corrupt autoencoders with deep regression networks, a deep alignment model robust to partial occlusions is achieved. Besides, our method can localize occluded regions rather than merely predict whether the landmarks are occluded. Experiments on two challenging occluded face datasets demonstrate that our method significantly outperforms the state-of-the-art methods.

1. Introduction

Face alignment or facial landmark detection, as an fundamental problem in computer vision, is widely used in face recognition, 3D face modeling, face animation, etc. For decades, many efforts are devoted to exploring robust alignment methods and great progresses have been achieved on face alignment under control condition or even in the wild [8, 7, 41, 34, 6, 29, 28, 40, 35, 11, 18, 32, 30, 27]. However, it is still a challenging problem due to various variations in appearance and shape, *e.g.*, pose, expression, especially partial occlusions. The alignment system usually degenerates severely under partial occlusions, and this problem is hard to tackle as any part of the face can be occluded by arbitrary objects.

Typical alignment models like Active Shape Models (ASMs) [8, 14] and Active Appearance Models (AAMs) [7, 23] employ Principal Component Analysis (PCA) to build a shape model or simultaneously establish shape and appearance models. These methods are sensitive to partial occlusions although the shape deformations are restricted by a PCA-based shape model.

Regression based methods have achieved impressive results on face alignment in the wild [11, 34, 2, 6, 24, 18, 28, 39]. Different from the parametric models ASM and AAM, these methods directly model the mapping from local features to the face shape with linear [34, 24] or deep architecture [28, 38, 39]. Especially, deep models (*e.g.*, Convolutional Neural Networks, Auto-encoders and Restricted Boltzmann Machines) make great progresses benefited from their favorable ability of modeling nonlinearity. Sun et al. [28] conduct a deep convolutional neural network to regress facial points by taking the holistic face as input and then cascade several convolutional neural networks to further refine the detection results within each local patch, which is extracted around the current face shape. Zhang et al. [39] design deep auto-encoder networks with joint shape-indexed features to regress the face shape. These methods achieve promising results on LFPW [3] and HELEN [20]. Benefited from the local features, these methods are somewhat robust to occlusions. However, they may still fail when severe partial occlusion occurs as they do not exclusively consider the occlusion problem.

Although the above methods have a certain tolerance of partial occlusion, the occlusion problem is not essentially solved. In this work, we propose a novel Deep Regression network coupled with De-corrupt Autoencoders (DRDA) to explicitly tackle the partial occlusion problem. As illustrat-

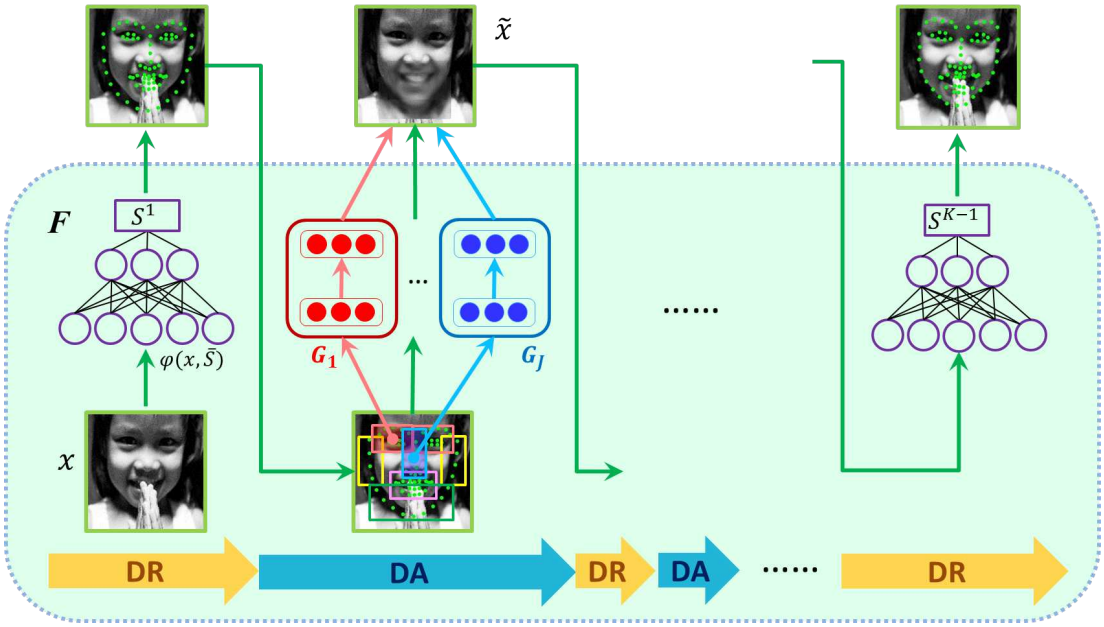

Figure 1. Overview of our DRDA for occlusion-free face alignment. F denotes the deep regression network which characterize the nonliner mapping from shape-indexed feature $\varphi(x, \bar{S})$ to a face shape, x is an input face and $\bar{S}$ is the initial face shape. G_i denotes the de-corrupt autoencoder for recovering from the occluded parts to achieve de-corrupted face image $\tilde{x}$, which can be leveraged in the following deep alignment model. Several stages are cascaded for both deep regression networks and de-corrupt autoencoders. DR is the abbreviation of deep regression networks and DA is the abbreviation of de-corrupt autoencoders.

ed in Figure 1, several deep regression networks are cascaded to characterize the complex nonlinear mapping from appearance to shape. The de-corrupt autoencoders are exploited to recover the occluded parts automatically. To recover a face with more appearance details under various poses and expressions, we partition the face into several components and establish one de-corrupt autoencoder for each of them. The de-corrupted image is then fed to the following deep regression networks for alignment task. By jointly learning de-corrupt autoencoders and deep regression networks, a deep alignment model robust to partial occlusions is attained. Moreover, different from previous works which focus on predicting whether the landmarks are occluded, our method can localize occluded regions rather than merely predict whether the landmarks are occluded.

The contributions are summarized as follows:

1. We present a novel face alignment method which can explicitly handle partial occlusion problem.
2. Different from the previous works that can only detect occlusions and discard the occluded parts, our proposed de-corrupt autoencoder network can automatically recover the occluded parts and the recovered parts can be leveraged together with those non-occluded parts, which leads to better alignment results.
3. Our method achieves the state-of-the-art results on the challenging datasets OCFW and COFW.

2. Related Works

Partial occlusions is a great challenge for face alignment. To improve the robustness to occlusions, many researches focus on implicitly or explicitly handling occlusions problem.

Most existing alignment methods achieve robustness to occlusions either by introducing shape constrains [8, 14, 17] or improving appearance representations [11, 34, 2, 6, 24, 38, 15, 21, 26, 9, 16]. For the first category, shape models are usually built to restrict the shape deformations so as to relieve the influence of occlusion. Huang et al. [17] integrate markov network inducing shape constrains to improve the robustness to occlusion. For the other category, part-based appearance representations are widely used to improve the robustness to occlusion. CLM [9] performs PCA on the concatenated local patches to build the appearance model. SDM [34] concatenates SIFT features extracted around landmarks and directly learns linear mapping from SIFT features to shape deviations. Differently, LBF [24] learns the linear mapping based on learnt local binary features of each facial point rather than the handcrafted SIFT, which achieves better performance than SDM. Local features have some tolerance to occlusion and the concentrated features can support each other to get reasonable results even if some patterns are occluded. Besides, Hu et al. [15] employ an active wavelet network (AWN) representation to replace the PCA-based appearance model

in AAM. Under occlusion, the reconstruction error spreads over the whole face in PCA-based appearance model while it remains local in wavelet network representation, leading to improved performance regarding partial occlusion. Liu et al. [21] design boosted appearance model (BAM) to substitute the appearance model in AAM and demonstrate the superiority for alignment under occlusion. To some extend, these methods may reduce the influence of occlusions. However, the occlusion problem is not settled in essence.

Recently, there are several methods attempting to explicitly handle occlusions problem [5, 33, 13, 37]. Burgos-Artizzu et al. [5] present a Robust Cascaded Pose Regression (RCPR) method to predict the landmarks as well as the occlusions. The face is divided into 3×3 grid and for each time, only one non-occluded block is utilized to predict the landmarks. It achieves impressive results on challenging dataset COFW which contains various occluded faces. However, the grid partition is not flexible and the features used for each regressor are limited within one block. Besides, training RCPR models needs occlusion annotations, which are expensive to achieve. Yu et al. [37] propose a occlusion-robust method recorded as Cor, which uses a Bayesian model to fuse the prediction results from multiple occlusion-specific regressors. Each regressor is trained by omitting some pre-defined parts in a face to tackle one type of occlusions. However, it is hard or even impossible to define occlusion-specific regressors to cover all types of occlusions considering the combinational explosion problem. Moreover, it may be time-consuming as there are several regressors for predicting a face shape. Golnaz et al. [13] explicitly model occlusions of parts into a hierarchical deformable part model to detect facial points and occlusions simultaneously. If a keypoint is occluded, the corresponding feature is discarded to avoid the influence of occlusions. It achieves better results than RCPR on COFW dataset. However it suffers from computational problem (about 30 seconds per image). Xing et al. [33] jointly learn an occlusion dictionary within a rational dictionary to capture different kinds of appearance under partial occlusions and get promising results on the occluded face dataset OCFW.

Overall, all these methods explicitly process occlusion problem either by discarding the occluded parts [5, 13, 37] or building an occlusion dictionary to capture the appearance variations due to occlusions [33]. Different from the existing methods, we explicitly tackle the occlusion problem by automatically recovering the occluded parts and then the de-corrupted parts are leveraged together with non-occluded parts for robust face alignment under occlusions.

3. Our Approach

In this section, we will firstly illustrate the overview of the proposed DRDA. Then we will demonstrate the details about deep regression networks for face alignment and de-corrupt autoencoders for recovering occluded face separately, followed by the learning of deep regression networks coupled with de-corrupt autoencoders under a cascade structure. Finally we will give a discussion about the differences with existing works.

3.1. Method Overview

This work attempts to learn a deep alignment model that is robust to partial occlusions. To this end, we propose a schema of coupling deep regression networks with de-corrupt autoencoders as shown in Figure 1. Firstly a deep regression network is employed to predict a face shape. After that, the de-corrupt autoencoders are designed to recover the occluded appearance around current shape. Then the de-corrupted face is taken as the input of successive deep regression network to further refine the face shape. As the occluded appearance is recovered in the de-corrupted face, more informative context rather than the worthless occlusion can be leveraged, leading to a better regression network for alignment. Furthermore, both deep regression networks and de-corrupt autoencoders can be conducted under a cascade structure, so as to attain a better and better de-corrupted face and face shape.

Suppose we have a training set $\{(x_i, S_i)\}_{i=1}^N$, which consists of N face images x_i and its corresponding p facial landmarks $S_i \in \mathbf{R}^{2p}$. The deep regression network is a nonlinear model that maps the image x_i to its corresponding face shape S_i by optimization the following objective:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \sum_{i=1}^{N} ||\Delta S_i - \mathbf{F}(\varphi(x_i, \bar{S}))||_2^2, \tag{1}$$

$$\Delta S_i = S_i - \bar{S}, \tag{2}$$

where φ is a feature extraction function, and $\bar{S} \in \mathbf{R}^{2p}$ is an initial shape. ΔS_i denotes the shape deviation between the groundtruth S_i and the initial shape $\bar{S}$.

In real-world scenarios, the input image x_i may be corrupted by occlusions. The corrupted features $\varphi(x_i, \bar{S})$ will significantly impact the learning of deep regression networks. As stated in [31], the corrupted parts can be recovered from partial reservations for images with redundance, such as face image. Inspired by this characteristic, we propose a de-corrupt autoencoder $\mathbf{G}$ to repair the destroyed patterns, which can be formulated as follows:

$$\mathbf{G}^* = \arg\min_{\mathbf{G}} \sum_{i=1}^{N} ||\hat{x}_i - \mathbf{G}(x_i)||_2^2. \tag{3}$$

Here, x_i is a corrupted face and $\hat{x}_i$ denotes the "pure" images without occlusions. Considering the the powerful recovering ability of autoencoders, $\mathbf{G}$ is modeled as an autoencoder-like neural network, denoted as De-corrupt Autoencoder.

With the "repaired" faces, the following deep regression networks can be relieved from suffering of occlusions. In the following, we will illustrate the deep regression networks $\mathbf{F}$ for shape prediction and the de-corrupt autoencoders neural networks $\mathbf{G}$ for recovering the occluded parts.

3.2. Deep Regression Network for Shape Prediction

The deep regression network $\mathbf{F}$ aims at characterizing the nonlinear mapping from appearance to shape. For a deep network with $m-1$ hidden layers, it can be formulated as optimizing the following objective function:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \sum_{i=1}^{N} ||\Delta S_i - f_m(f_{m-1}(...f_1(\varphi(x_i, \bar{S})))||_2^2 + \lambda \sum_{i=1}^{m} ||W_i||_F^2, \tag{4}$$

$$a_q \triangleq f_q(a_{q-1}) = \sigma(W_q a_{q-1} + b_q), q \in [1, m-1], \tag{5}$$

$$f_m(a_{m-1}) = W_m a_{m-1} + b_m, \tag{6}$$

where $\mathbf{F}$ consists of $f_1, f_2, ..., f_m$. f_q is the nonlinear function of qth layer in the deep network parameterized with W_q and b_q, $q \in [1, m-1]$. $\varphi(x_i, \bar{S})$ denotes the shape-indexed SIFT [22] features extracted around the shape $\bar{S}$. σ is the nonlinear activation functions, *e.g.*, sigmoid or tanh function. a_q denotes the response of hidden layer q. For the last layer m, a linear regression is utilized to project the feature representation a_{m-1} to the corresponding face shape deviation ΔS_i. A regularization term $\sum_{i=1}^{m} ||W_i||_F^2$ is introduced to prevent over-fitting. Eq. (4) can be optimized by using the algorithm of L-BFGS [19].

3.3. De-corrupt Autoencoders for Recovering the Occluded Face

Usually, the deep regression model can achieve an accurate face shape. However, under the presence of partial occlusions, the performance may suffer from degenerating. To tackle the partial occlusions, we design an autoencoder-like neural network to explicitly recover the occluded parts, denoted as De-corrupt Autoencoder.

Auto-Encoder. A conventional auto-encoder network consists of two components, *i.e.*, encoder and decoder [4]. It is trained by minimizing the reconstruction error of the input, which is purely unsupervised. The encoder function g is exploited to map the input vector $x \in \mathbf{R}^d$ to a hidden representation $y \in \mathbf{R}^r$, where r is the number of hidden units. The mapping function g is illustrated as follows:

$$y = g(x) = \sigma(Wx + b), \tag{7}$$

where W is a $r \times d$ weight matrix and $b \in \mathbf{R}^{r \times 1}$ is a bias term. σ is a sigmoid function or tanh function, which induces the nonlinearity. The decoder function h attempts to map the representation y back to x:

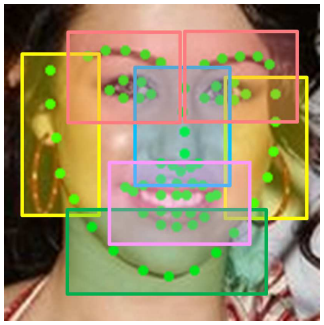

Figure 2. Component partition of 68 facial points. The 68 facial landmarks are divided into seven components, *i.e.*, left-eye-eyebrow, right-eye-eyebrow, nose, mouth, left-face-contour, right-face-contour and bottom-face-contour.

$$h(y) = \sigma(W'y + b'), \tag{8}$$

where $W' \in \mathbf{R}^{d \times r}$, $b' \in \mathbf{R}^{d \times 1}$. W' is optionally be constrained by $W' = W^T$, in which it calls as tied weights. Given a training set with N samples $\{x_i\}_{i=1}^N$, the parameters of encoder and decoder functions are optimized by minimizing the reconstruction error as follows:

$$\{W^*, b^*, W'^*, b'^*\} = \arg\min_{W,b,W',b'} \sum_{i=1}^{N} ||x_i - h(g(x_i))||_2^2, \tag{9}$$

De-corrupt Autoencoder. Inspired by the favorable ability of autoencoder for reconstruction, the proposed de-corrupt autoencoder is designed to recover the genuine appearance from the occlusions. Assume we have a training set with N samples $\{(x_i, \hat{x}_i, \hat{S}_i)\}_{i=1}^N$, x_i is a occluded face with the shape $\hat{S}_i$ and $\hat{x}_i$ is the corresponding genuine face without occlusions. Here, $\hat{S}_i$ is the predicted shape from the anterior deep regression network. A de-corrupt autoencoder network $\mathbf{G}$ is designed to reconstruct the genuine face $\hat{x}_i$ from the occluded one x_i as below:

$$\{W^*, b^*, W'^*, b'^*\} = \arg\min_{W,b,W',b'} \sum_{i=1}^{N} ||\hat{x}_i(\hat{S}_i) - h(g(x_i(\hat{S}_i)))||_2^2, \tag{10}$$

where $x_i(\hat{S}_i)$ denotes the appearance of x_i around the shape $\hat{S}_i$, and $\hat{x}_i(\hat{S}_i)$ denotes the appearance of $\hat{x}_i$ around the shape $\hat{S}_i$. Generally, the encoder function g and deocoder function h can be designed with multiple layers.

Considering that the face appearance varies under different poses and expressions, it is nontrivial to design one de-corrupt autoencoder network to reconstruct the details of the whole face. To recover a face with vivid appearance under various poses and expressions, we partition the face image

x_i into J components according to face shape $\hat{S}_i$ and design several independent de-corrupt autoencoder networks $\mathbf{G}_j$, one for each component. Each of the J components is denoted as $x_i(\hat{S_{ij}})$ with $\hat{S_{ij}}$ representing those facial points belonging to the jth component, $j \in [1, J]$. As shown in Figure 2, 68 facial points are partitioned into seven components. The region size of each component is calculated as the rectangular hull around the relevant facial points.

To attain de-corrupt autoencoder networks, we need to build a training set consisting of x_i and the genuine version $\hat{x}_i$. However, occluders have thousands of appearance variations and occlusion may occur anywhere. It is difficult or even impossible to collect real-world images with a variety of possible occlusions. Fortunately, it is easy to collect images $\hat{x}_i$ without occlusions. And a large scale of face images with occlusions can be obtained by randomly putting occluders on it. Specifically, for each component j, we randomly pick patches from nature images which contain no face, and then randomly place the patch anywhere within this component, inducing an occluded face component.

With the de-corrupt autoencoder network $\mathbf{G}$, a corrupted face image x_i can be recovered as $\mathbf{G}(x_i)$. Instead of directly using the recovered $\mathbf{G}(x_i)$, we only leverage the recovered appearance for those occluded parts, but original appearance for those non-occluded parts, forming the de-corrupted image $\tilde{x}_i$ used for further refining the shape. Specifically, the occluded parts are determined as follows: firstly, compute the difference between x_i and $\mathbf{G}(x_i)$; then simply regard those pixels with the difference larger than a threshold τ (τ=30 in this work) as occluded pixels; finally those components with the proportion of occluded pixels more than 30% are determined as occluded parts.

3.4. Cascade Deep Regression with De-corrupt Autoencoder

Face shape S^1 can be attained from the network $\mathbf{F}$ stated above. However, it may not approach the ground truth due to partial occlusion, pose, expression, *etc*. To achieve finer alignment results, several successive deep regression networks coupled with de-corrupt autoencoders are cascaded. Specifically, for the kth stage, the de-corrupt autoencoder network $\mathbf{G}_j^k$ is constructed for the jth component based on current shape prediction S^{k-1}, formulated as follows:

$$\mathbf{G}_j^{k*} = \arg\min_{\mathbf{G}_j^k} \sum_{i=1}^{N} ||\hat{x}_i(S_{ij}^{k-1}) - h_j^k(g_j^k(x_i(S_{ij}^{k-1})))||_2^2. \tag{11}$$

With $\mathbf{G}_j^{k*}|_{j=1}^J$, a de-corrupted face $\tilde{x}_i^{k-1}$ is achieved. By taking $\tilde{x}_i^{k-1}$ as input, the following deep regression network of stage k is designed to further refine the shape by predicting the current shape deviation $\Delta S^k = S - S^{k-1}$:

$$\mathbf{F}^{k*} = \arg\min_{\mathbf{F}^k} \sum_{i=1}^{N} ||f_m^k(f_{m-1}^k(...f_1^k(\varphi(\tilde{x}_i^{k-1}, S_i^{k-1})))) - \Delta S_i^k||_2^2 + \lambda \sum\nolimits_{i=1}^{m} ||W_i^k||_F^2, \tag{12}$$

where $\varphi(\tilde{x}_i^{k-1}, S_i^{k-1})$ denotes the shape-indexed feature extracted from the de-corrupted face image $\tilde{x}_i^{k-1}$ with the current shape S^{k-1}. With Eq. (12), the shape can be refined as $S^k = S^{k-1} + \Delta S^k$ which is further utilized for learning the following de-corrupt autoencoder networks.

By learning de-corrupt autoencoder networks and deep regression networks under a cascade structure, they can benefit from each other. On the one hand, with more accurate face shape, the appearance variations within each component becomes more consistent, leading to more compact de-corrupt autoencoder networks for better de-corrupted face images. On the other hand, the deep regression networks that are robust to occlusions can be attained by leveraging better de-corrupted faces. The final shape prediction S^K can be achieved by summing the outputs of each stage: $S^K = \bar{S} + \sum_{k=1}^{K} \Delta S^k, \Delta S^k = \mathbf{F}^k(\varphi(\tilde{x}^{k-1}, S^{k-1}))$. Besides, the occluded regions (*i.e*., the occluders) can be localized by comparing the difference between the final de-corrupted image $\tilde{x}^{K-1}$ and the origin x.

3.5. Discussions

In this section, we give a brief discussion about the differences between our method and some existing methods which also explicitly tackle occlusions.

Differences from [5, 13, 37]. All these methods [5, 13, 37] handle partial occlusion problem by learning occlusion-aware face alignment model, *i.e*., explicitly considering the occlusion state when building alignment models, as well as ours. However our method differ from them in two aspects: 1) [5, 13, 37] discard the occluded parts to achieve the face alignment model robust to occlusions, while our proposed de-corrupt autoencoder network automatically recovers the occluded parts which can be further utilized together with those non-occluded parts for predicting face shape, leading to better alignment results. 2) The existing methods can only predict whether the landmarks are occluded while ours can easily localize the exact occluded regions with the de-corrupt autoencoder as shown in Figure 5, see details in Sec. 4.3.

Differences from [33]. Xing et al. [33] propose a novel method named OSRD to deal with occlusions by learning an occlusion dictionary and a rational dictionary. However, it may be hard for a linear model to well capture more complex variations. In contrast, our de-corrupt autoencoder networks can well undo the challenging corruption, attributed to the favorable ability for modeling the nonlinearity.

4. Experiments

We firstly introduce the evaluation settings including the datasets and methods for comparison, and then illustrate the implementation details of our method. Finally, we compare the proposed approach with the state-of-the-art methods on three challenging datasets under various occlusions, *i.e*., OCFW [33], COFW [5] and IBUG [25].

4.1. Datasets and Methods for Comparison

To evaluate the effectiveness of the proposed method, we employ two challenging datasets with varied occluded faces: **OCFW** [33] and **COFW** [5]. OCFW contains 2591 images for training and 1246 images for testing, which comes from **LFPW** [3], **HELEN** [20], **AFW** [41] and **IBUG** [25]. The images are collected in the wild, which contain large variations in pose, expression, partial occlusion, *etc*. All training samples are without occlusions while all testing samples are partially occluded, which formulates a challenging scenario of face alignment under occlusions. Annotations of 68 landmarks are published in website [1]. COFW is another occluded face dataset in the wild, which is published by Burgos-Artizzu et al. [5]. Its training set consists of 845 faces from LFPW training set and extra 500 faces with heavy occlusions. The test set contains 507 faces which are also heavily occluded. The faces have large variations in varying degrees of occlusions, together with different poses and facial expressions. Both 29 landmarks and their occluded/unoccluded state are released in [5]. Besides, we also do comparisons on 135 images from IBUG dataset which formulates a more general "in the wild" scenario.

We compare our method with a few state-of-the-art, *i.e*., SDM [34], RCPR [5], OSRD [33]. For RCPR, we use the off-the-shelf codes released by the original authors. Since the released SDM model only predicts the inner 49 landmarks, we re-train SDM to predict 68 landmarks. For OSRD, we quote results from [33]. We cascade several deep regression networks without de-corrupt autoencoders as a baseline, denoted as CDRN. The normalized root mean squared error (NRMSE) is calculated with respect to the interocular distance. The cumulative distribution function (CDF) of the normalized root mean squared error (NRMSE) is employed for performance evaluation and the NRMSE is calculated with respect to interocular distance.

4.2. Implementation Details

Our proposed deep alignment model consists of 3 stages. For the first stage, the regression network has three layers including two non-linear hidden layers and one linear layer. The numbers of hidden neurons are set as 784 and 400 respectively. For the following two stages, the regression network is deeper, which contains three hidden layers with 1296, 784 and 400 hidden neurons respectively, and one linear regression layer. The network of the first stage takes face of 80×80 as input and for the last two stages, face image of higher resolution (*i.e*., 160×160) is used for finer shape prediction. The initial shape $\bar{S}$ is a mean shape based on face detection results. For all stages, The weight decay parameter α is set as 0.001, which balances the sum-of-squares error term and the weight decay term. We divide the face shape of 68 landmarks into seven components, and for each component, a de-corrupt autoencoder network with one hidden layer is utilized to recover the occluded parts. The number of hidden units is set as 576 for all de-corrupt autoencoders. Sigmoid function is used to induce nonlinearity for both deep regression networks and de-corrupt autoencoders.

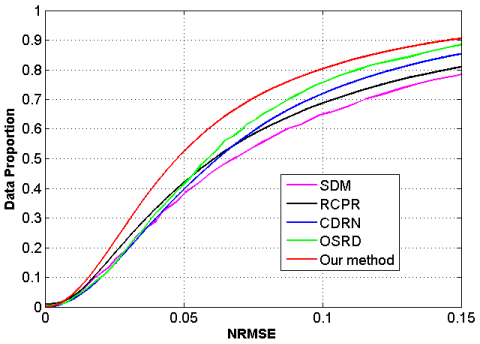

Figure 3. Evaluations on OCFW.

(a) RCPR Fitting Results

(b) CDRN Fitting Results

(c) The proposed DRDA Fitting Results

Figure 4. Visualized results of RCPR, CDRN and our proposed DRDA on OCFW. The first row shows the results of RCPR. The second row shows the results of CDRN and the results of our proposed DRDA are illustrated in the last row.

4.3. Evaluations on OCFW Dataset

Firstly, we evaluate our method and the existing approaches on OCFW dataset. All methods are trained on

Figure 5. Exemplar results of de-corrupted face images and occluded parts localizations on OCFW. The first row shows the origin occluded faces. The de-corrupted images are shown in the second row. The last row shows the occlusion localization results.

OCFW training set and then are evaluated on OCFW test set in terms of 68 facial landmarks.

Figure 3 shows the cumulative error distribution curves of all methods. As seen, RCPR performs better than SDM. It is possibly because RCPR designs the interpolated shape-indexed features which are robust to large pose, and applies smart restart strategy to relieve the sensitivity of shape initialization. Furthermore, CDRN performs slightly better than RCPR when the NRMSE is above 0.065, which can be attributed to the favorable ability of modeling nonlinear mapping from appearance to shape. Benefited from simultaneously learning an occlusion dictionary and a rational dictionary, OSRD achieves better results than both RCPR and CDRN. Furthermore, our method outperforms OSRD, with an improvement up to 10% when NRMSE is 0.05. This significant improvement can be attributed to the schema of effectively coupling deep alignment networks with de-corrupt autoencoders. Figure 4 shows the landmark detection results of RCPR, CDRN and our DRDA on some extremely challenging samples. It can be observed that our method improves the robustness of face alignment to partial occlusions under varied situations. Besides the face alignment, our proposed method can roughly localize the occluded region by comparing the de-corrupted image with the origin occluded image, as stated in Sec. 3.3. Some de-corrupted images and occlusion localization results are illustrated in Figure 5. As seen, our method can well recover the genuine appearance from the occlusions as well as localize the occluded parts.

4.4. Evaluations on COFW Dataset

We further evaluate the detection accuracy of all methods on Caltech Occluded Faces in the Wild (COFW). COFW is another challenging dataset with real-world faces occluded to different degrees. And it also contains large shape and appearance variations due to pose and expressions. All methods are evaluated on COFW test set in terms of 29

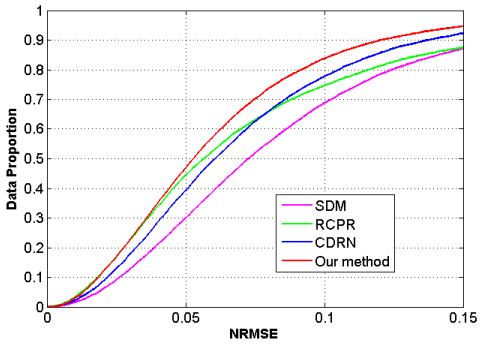

Figure 6. Evaluations on COFW.

facial landmarks according to the official protocol. For RCPR, we use the model released by authors for evaluations, which is trained with occlusion annotations for robust face alignment under occlusions and achieves promising results on COFW test set [5]. For the rest methods, including ours, we directly use the model trained on OCFW training set for evaluation. Since the model trained on OCFW training set is to predict 68 landmarks, we follow [13] to learn a linear mapping from the predictions to the 29 facial points.

The performances of all methods are illustrated in Figure 6. Simliar conculusions can be achieved. As seen, RCPR performs better than SDM and CDRN when the NRMSE is lower than 0.08. RCPR explicitly predicts occlusions and utilizes the occlusion information to help shape prediction, so as to achieve robustness to partial occlusions. Benefited from the de-corrup autoencoder networks, Our method performs better than RCPR and CDRN, with an improvement up to 6% when NRMSE is 0.10. Besides, we compare our method with methods [13, 12] in terms of the mean error. In [13, 12], they achieves $7.46(\times 10^{-2})$ and $7.30(\times 10^{-2})$ respectively on COFW testset and ours achieves a much better results of $6.46(\times 10^{-2})$, which also demonstrates the superiority of our effective alignment framework for dealing with occlusions. Some alignment results of our method are shown in Figure 7. The first row shows the alignment results under occlusions simultaneously with varying poses. The second row exhibits exemplars under simultaneous occlusions and expressions. Samples with a variety of occluders (*e.g.*, sunglasses, respirator, cameras, *etc.*) are illustrated in the last row. As seen, our method is robust to different types of occlusions under various poses and expressions.

Since COFW consists of occlusion annotations, we do quantitative evaluation of occlusion detection on COFW and compare with RCPR [5], OC [13], CoR [37]. The occlusion detection precision/recall curves are shown in Fig. 8. As can be seen, our method significantly outperforms the others. The visualized results of occlusion detection are shown in Fig. 9. Our method can not only detect occlusions accurately but also well recover occluded parts.

Figure 7. Exemplar results on COFW. The first row: occluded images under various poses; the seconde row: images under occlusions together with different expressions; the last row: images occluded by a variety of objects.

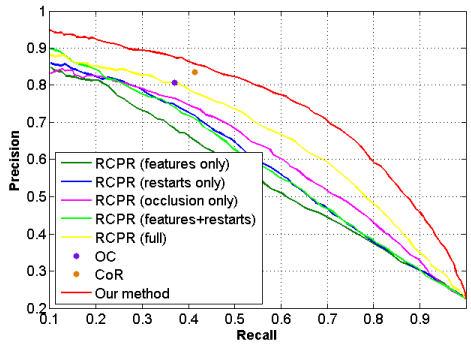

Figure 8. Occlusion detection precision/recall curves on COFW.

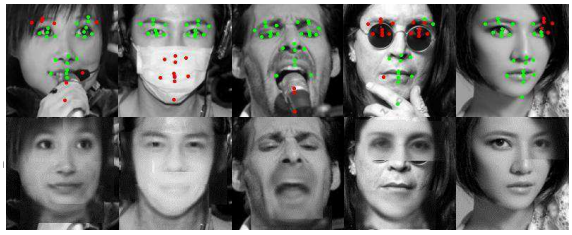

Figure 9. Visualized results of occlusion detection and reconstruction on COFW. The first row shows the occlusion detection results, where the red dots denotes occlusion and the green dosts denotes non-occlusion. The corresponding de-corrupted faces are shown in the last row.

4.5. Evaluations on IBUG Dataset

Besides, we conduct experiments on IBUG dataset, which formulates a more general scenario containing large variations due to pose, expression, occlusion, illumination, *etc*. We compare our method with more state-of-the-art works, including Dantone et al. [10], Zhu et al. [41], Yu et al. [36], DRMF [2], SDM [34], RCPR [5]. All methods are trained with LFPW trainset, HELEN trainset and AFW and evaluated on IBUG dataset in terms of 68 landmarks.

Fig. 10 shows the comparison results of all methods

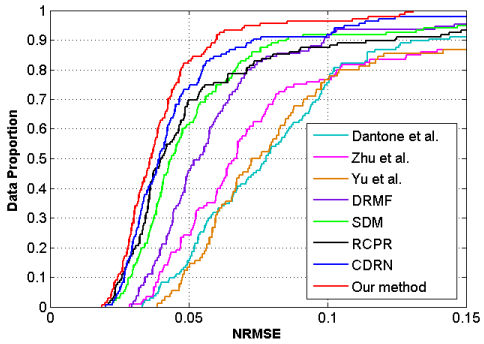

Figure 10. Evaluations on IBUG.

with respect to cumulative error distribution curves. The NRMSE is normalized by the face size for better exhibition. As shown in Fig. 10, our method achieves better results than the state-of-the-art methods and also outperforms the CDRN, which demonstrates the effectiveness of inducing de-corrupt autoencoder networks. Besides, we compare ours with LBF [24]. The mean error of LBF is 11.98% while our method achieves a better result of 10.79%.

5. Conclusions and Future Works

In this work, we present a novel Deep Regression networks coupled with De-corrupt Autoencoders (DRDA) for face alignment under partial occlusions. Aiming at explicitly tackling the occlusion problem, we design the de-corrupt autoencoder networks to automatically recover the genuine appearance for the occluded parts and leverage the recovered parts together with the non-occluded parts in the deep regression network to get an accurate shape prediction under occlusions. By learning the de-corrupt autoencoders and deep alignment networks under a cascade architecture, the de-corrupt anutoencoder network becomes more and more compact with better shape predictions, and the alignment model becomes more and more robust with improved de-corrupted faces, leading to accurate face alignment results. Our method achieves the state-of-the-art performance on three challenging datasets consisting of various occlusions together with large pose and expression variations, which demonstrates the effectiveness of our DRDA for face alignment under occlusions. In the future, we will explore other types of deep architectures for recovering the genuine appearance for occluded parts.

Acknowledgements

This work was partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61402443, 61222211, 61272321, and the Strategic Priority Research Program of the CAS (Grant XDB02070004).

References

[1] 300 faces in-the-wild challenge. http://ibug.doc.ic.ac.uk/resources/300-W/.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[4] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2009.

[5] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[6] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2001.

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding (CVIU)*, 1995.

[9] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *The British Machine Vision Conference (BMVC)*, 2006.

[10] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2578–2585, 2012.

[11] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[12] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, and X.-J. Wu. Random cascaded-regression copse for robust facial landmark detection. *IEEE Signal Processing Letters*, 2015.

[13] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[14] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *European Conference on Computer Vision (ECCV)*. 2008.

[15] C. Hu, R. Feris, and M. Turk. Active wavelet networks for face alignment. In *The British Machine Vision Conference (BMVC)*, 2003.

[16] C. Hu, R. Feris, and M. Turk. Real-time view-based face alignment using active wavelet networks. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFGW)*, 2003.

[17] Y. Huang, Q. Liu, and D. N. Metaxas. A component-based framework for generalized face alignment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2011.

[18] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] Q. V. Le, A. Coates, B. Prochnow, and A. Y. Ng. On optimization methods for deep learning. In *International Conference on Machine Learning (ICML)*, 2011.

[20] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision (ECCV)*. 2012.

[21] X. Liu. Generic face alignment using boosted appearance model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004.

[23] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 2004.

[24] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[25] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013.

[26] J. M. Saragih, S. Lucey, and J. F. Cohn. Face alignment through subspace constrained mean-shifts. In *The IEEE International Conference on Computer Vision (ICCV)*, 2009.

[27] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[28] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[29] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *The IEEE International Conference on Computer Vision (ICCV)*, 2013.

[30] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[31] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning (ICML)*, 2008.

[32] Y. Wu, Z. Wang, and Q. Ji. 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)*, 2013.

[33] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan. Towards multi-view and partially-occluded face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[34] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[35] X. Xiong and F. De la Torre. Global supervised descent method. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[36] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[37] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *European Conference on Computer Vision (ECCV)*. 2014.

[38] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision (ECCV)*. 2014.

[39] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*. 2014.

[40] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[41] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.