ROBUST FACE ALIGNMENT WITH CASCADED COARSE-TO-FINE AUTO-ENCODER NETWORK

Cheng Peng, Yongxin Ge, Mingjian Hong, Sheng Huang, Dan Yang*

School of Software Engineering, Chongqing University, Chongqing, China
*Email: yongxinge@cqu.edu.cn

ABSTRACT

In this paper, we present a novel face alignment method using a two-level cascaded auto-encoder networks (2-LCAN). In our framework, the first level auto-encoder networks generate rough facial landmarks locations by taking detected face images with low-resolution as inputs. The second level auto-encoder networks are constructed by cascading several sub stacked auto-encoder networks (SSAN) in a coarse-to-fine manner. Each SSAN extracts SIFT features and local pixels features around current landmark positions, then fuses them together to further refine landmarks of different facial components with higher image resolutions. Finally, experimental results on LFPW and HELEN datasets demonstrate that our proposed method is significantly superior to the compared approaches both in accuracy and robustness.

Index Terms— Face Alignment, Deep Learning, Auto-encoder Network, Coarse-to-Fine

1. INTRODUCTION

Face alignment or facial landmark detection aims at automatically and precisely locating facial landmark points, such as eyebrow, nose, eye, mouth and cheek contours, in a holistic face image. As a crucial step in face recognition [1], facial expression synthesis [2] and age estimation [3], face alignment has become a popular research point in computer vision tasks over the past few years. Although many current face alignment works can localize landmarks accurately, it is still difficult in addressing face alignment issue in wild or in other extreme cases that caused by the large pose variations and occlusions.

1.1. Related Work

Over decades, numbers of face alignment methods have been proposed, many of which are proved to be efficient and work quite well in some specific applications. Overviewing the whole development of the face alignment approaches, they can be approximately divided into three categories: discriminative fitting [4, 5, 6, 7], shape regression [8, 9, 10, 11] and deep learning [12-17].

Classic discriminative fitting approaches such as Active Shape Models (ASMs) [4] and Active Appearance Models (AAMs) [5] apply Principal Component Analysis (PCA) to create a parametric model between facial shape and texture that is based on training dataset. Constrained Local Models (CLMs) [7] also employ mean facial shape as initialization, and then create an appearance model based on the local patches around these landmarks. However, all above methods suffer in partial occlusions and hardly generate a robust enough model.

Shape regression approaches are widely used for accurate face alignment. Differing from discriminative fitting approaches, regression based methods are given an initial shape S_0 (usually the mean facial shape [10]), and then directly estimate a shape incremental deviation ΔS by fitting a feature-to-shape regression. For example, Burgos-Artizzu et al. [9] propose a Robust Cascaded Pose Regression model based on [8]. They initialize the facial shape several times to get a more robust initialization. By introducing occlusion information, they train a series of boosted regressors and leverage random ferns to update the shape deviation for landmark detection. Xiong et al. [11] present a supervised descent method (SDM) for face alignment. They gradually map feature-to-pose by performing cascaded regressions that are based on SIFT features [18]. While these shape regression methods have achieved promising performance, most of them use mean facial shape as initialization, which greatly affect the results of subsequent regressions. Furthermore, simple linear regression cannot solve the complex non-linear mapping from extracted features to face shape ideally.

Due to the excellent performance of deep neural networks on working out non-linear problems, deep learning approaches are increasingly used in face alignment. In [12], Sun et al. proposed a three-stage framework, where Deep Convolution Neural Network (DCNN) were used to regress local landmarks independently. Instead of leveraging DCNN, Zhang et al. [14] cascaded four stacked auto-encoder networks (SAN) for real-time face alignment. However, their three local SANs concatenate all local shape-indexed features together for regression, which may ignore the impacts of occlusions that in local pixel patch.

In our work, we proposed a robust face alignment method called two-level cascaded auto-encoder networks (2-LCAN). As briefly illustrated in Fig. 1, the architecture of our 2-LCAN is consisted of one SAN and three SSANs. The first LCAN generates an initial shape which is insufficiently

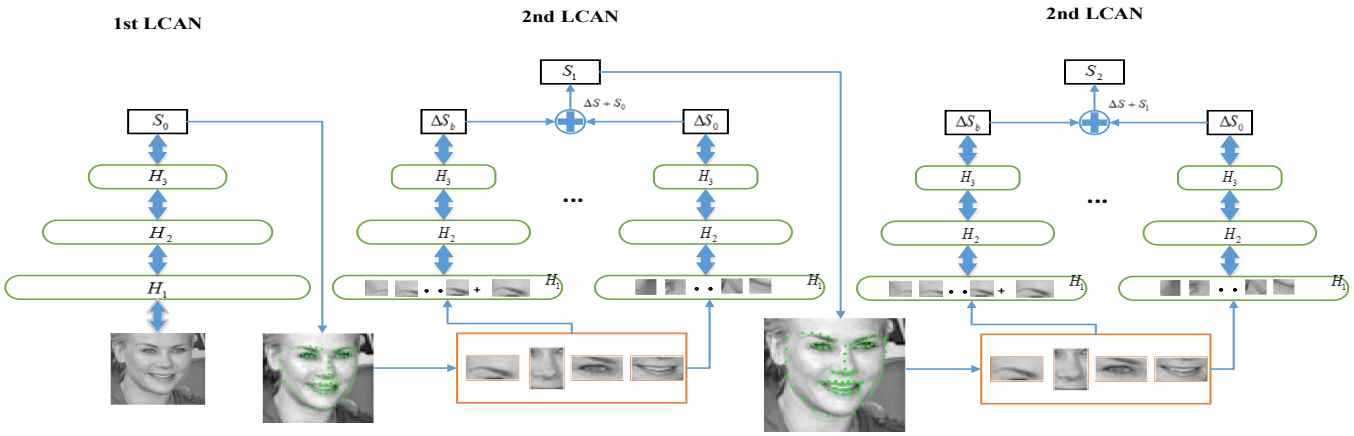

Fig.1. Overview of our 2-LCAN for robust facial landmark detection. The 1st LCAN quickly generates an insufficiently accurate facial shape. The 2nd LCAN further refine facial components separately as the image's resolutions are increased higher and higher. Due to the format of paper, we miss the last LCAN in Fig.1.

accurate. Then the second LCAN divides the entire face into five components (eyebrow, nose, eye, mouth and cheek contours) according to the current shape. Finally, we extract features in the five local parts with gradually higher resolution and use them to fit the non-linear feature-to-shape regression for further refinement. The experimental results demonstrate the outstanding performance of our approach.

The main contributions are summarized as follows:

1) We cropped the entire face according to eyebrow, nose, eye, mouth and cheek contours in the second LCAN, which can separate influences of each loss functions and balance the difficulties of localizing different landmark positions.
2) We extracted local pixels features as well as local SIFT features, which may weaken the impact of occlusions in regression.
3) We proposed a smart training method that divided our training data into four parts for different levels' training to simulate the testing environment.

2. TWO-LEVEL CASCADED AUTO-ENCODER NETWORK

2.1. First LCAN

In face alignment, a good initial estimation of the face shape tends to be able to get a more accurate feature-to-shape result. Most previous regression fitting methods [5, 11] use mean facial shape (calculate from the training set) as an initialization. However, this casual initialization may lead to the situation that the regression model fall into local minimum. Inspired by [14, 15], we also leverage a stacked auto-encoder networks as our first LCAN to predict a more robust initial shape.

Assuming $X = [x_1, x_2, \dots, x_i]$ are the low-resolution images of the training set, where x_i represents the pixel vector of the ith image. $S_g = [x_1, y_1; \dots; x_p, y_p]$ denote the ground truth positions of p landmarks. Before training, all pixel values of images are normalized to [0, 1] for better compatible with the SAN. The mapping objective function F of pixel-to-shape can be formulated as follows:

$$F = \arg\min_{l,f} \sum_{x_i \in X} \left\| S_g - l(f_k(x_i)) \right\|_2^2 + \alpha \sum \|W\|_2^2 \quad (1)$$

$$f_k = \sigma(W_k x_{k-1} + b_k) \quad (2)$$

$$h_k = \sigma(W_k^{\mathrm{T}} f_k + b'_k) \quad (3)$$

Where f_k is the encoding function used in three encoding layers, as shown in Eq. (2). It encodes x_i to global pixels features which can better represent the input x_i. l is a linear regression function in regression layer that maps the global pixels features to facial shape. $\alpha \sum \|W\|_2^2$ is a weight decay term to prevent over-fitting.

$$\arg\min_{W,b} \|x_i - h_k(f_k(x_i))\|_2^2 \quad (4)$$

For the training of the first LCAN, we first set parameters by training three encoding layers in an unsupervised greedy methods in Eq. (4), specifically, h_k is a decoding function. Then we initialize the parameters of linear regression layer randomly, and fine-tune the whole networks with Eq. (1) at last.

2.2. Second LCAN

The 1st LCAN has been able to generate an initial shape according to the different input facial images, while the accuracy of which is still far from satisfactory since the challenges of nonlinear mapping from feature to shape. To further refine the landmark positions, we proposed a novel coarse-to-fine framework, the Sub SAN (SSAN), and we cascaded several successive SSANs as our second LCAN. As illustrate in Fig. 1, there are five SSANs in each 2nd LCAN, and every SSAN has three encoding layers and one linear regression layer as well as the 1st LCAN. We train five

different SSANs to refine the ΔS of different facial components (eyebrow, nose, eye, mouth and cheek contours) separately rather than train one individual SAN for all of the facial landmarks. This is because the difficulty of localizing distinct landmarks are unbalanced. For example, the landmark detection of mouth and cheek are much more difficult to any other facial components. For mouth, there are numerous of actions that are associated with mouth in face datasets, such as singing, smiling, laughing and so on. All these pose variations in mouth may increase the difficulty of the alignment as well as occlusions. For cheek, there are less local texture information than other inner face components and the alignment of cheek contours is easily influenced by the noises from the background. That is to say, the alignment error of mouth and cheek may dominate the updates of weights in networks that are computed by loss function if all landmarks are trained together. Hence, we separate the single SAN to five SSANs, so that the different SSANs are able to refine landmark positions in their own loss function instead of sharing parameters with those 'difficult ones'. For example, the loss functions of eyebrow are defined as

$$\arg\min_{l,y} \left\| \Delta S_j^0 - l^0\left(y_k^0\left(\varphi\left(S_{j-1}^0\right)\right)\right)\right\|_2^2 \quad (5)$$

Where $\varphi(S_0)$ is the Local Shape-Indexed feature which has been widely used in many deep learning methods [16, 17, 19], and most of these previous methods achieve promising performance for face alignment. In our SSANs, we also leverage local SIFT features as inputs. After the 1st LCAN, we crop the holistic face image into five components (eyebrow, nose, eye, mouth and cheek) according to the current shape S_0, then we extract SIFT features in the local patches around the landmarks which are included in these five components and concatenate them into five feature vector separately. While the local SIFT features can offer us robust pose information for face alignment, it is also easily suffer from occlusion. As shown in Fig. 2, if large parts of the local patches are occluded, the local SIFT features which are extracted from these patches are useless and even interfere the alignment of landmarks.

Motivated by the above observation, we extract SIFT features and local pixel features simultaneously and fuse them together as the input of each SSAN. The local pixel features $\vartheta(S_{j-1}^{(n)})$ are encoded by an encoding layer from the nth local facial component. When the SIFT features are extracted from occluded patches, local pixel features can provide a global constraints in these local facial components by offering more complete texture information, which will better help SSAN get the prediction of ΔS. So the objective function Y can be redefined as follows:

$$Y = \arg\min_{l,y} \left\| \Delta S_j^n - l^n\left(y_k^n\left(\varphi\left(S_{j-1}^n\right), \vartheta\left(S_{j-1}^n\right)\right)\right)\right\|_2^2 + \alpha \sum \left\|W^j\right\|_2^2 \quad (6)$$

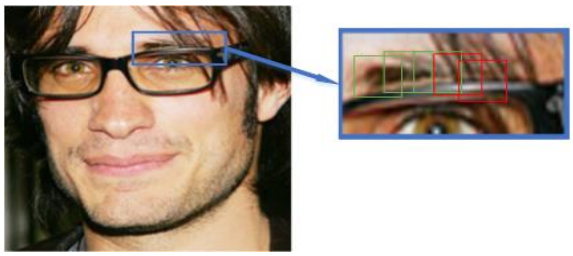

Fig.2. The right side is a detected face image. The blue patch indicates right eyebrow component. The left side is its enlarged view, and the five smaller patches show the local patches around the right eyebrow's landmark positions. Green patches represent few occlusions, while the red patches are mostly occluded, which are unreliable.

$$\Delta S_j = S_j - S_{j-1} \quad (7)$$

The first term of Eq. (5) is the new loss function, and the second term is a global weight decay term that decays the training parameters. We take the same pre-training strategy as the first LCAN does, and then fine-tune each SSANs with Eq. (5). After the whole SSANs have been trained and the facial shape update ΔS_j have been calculated, we replace S_j by $S_j = S_{j-1} + \Delta S_j$. Finally, we employ several second LCANs in a cascaded manner and repeat the above steps to refine the facial shape gradually.

2.3. Smart Training

Deep neural network is easy to be over-fitting. To solve this problem, we proposed a smart training method. Instead of training our model with whole training set, we divide the training set into three parts randomly before the start of training. Then we take the first part as our 1st LCAN's input and fix the trained parameters. The 1st LCAN is robust, but it may influence following networks' generalization ability. For this purpose, we predict S_0 with the previous trained parameters by inputting the second part of training set. At last, we extract features around S_0 to do the training of 2nd LCAN and repeat these above operations for following training. The experimental results demonstrate that our smart training can simulate the test situation well.

3. IMPLEMENTATION DETAILS

We choose 68 landmarks annotation [20, 21, 22] for each face image. Our 2-LCAN is constituted of one 1st LCAN and three 2nd LCANs. Both of which have three encoding layers and one linear regression layer as illustrated in Fig. 1. More specifically, there are 1300, 700, 300 hidden units in each layer of the 1st LCAN, and the numbers of hidden units for five SSANs are diverse from each other. In Eq. (1) and (6), all weight decay parameters α are set as the same value, $\alpha = 0.001$. The initial size of face images are uniformly set in low-resolution as 50×50 pixels, and the following resolutions in refinement steps are increasingly higher.

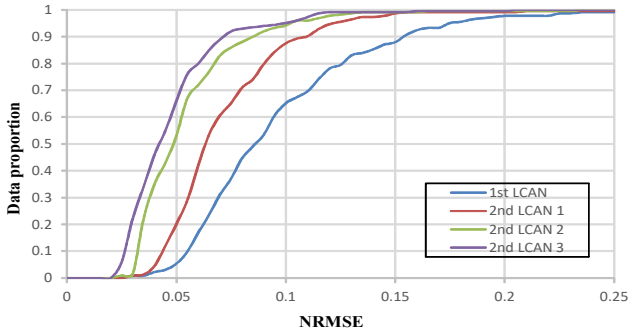

Fig.3. The cumulative error distribution (CED) curves from LFPW of 2-LCAN.

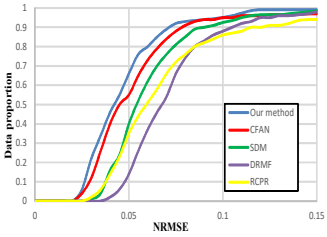

Fig.4. Comparison on LFPW

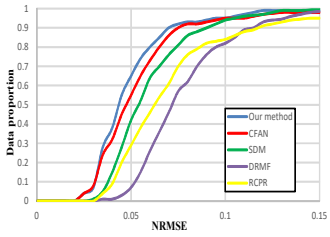

Fig.5. Comparison on HELEN

4. EXPERIMENTAL RESULTS

In this section, we first introduce the experimental datasets and settings. Then, we investigate the results of 2-LCAN separately. Finally, we compare our 2-LCAN with several state-of-the-art methods and investigate our improvements.

4.1. Experimental Datasets

1) LFPW [20]: There are 1432 facial images in the original LFPW dataset, while some URLs of the original LFPW images are invalid, there are only 811 training images and 214 testing images on the IBUG website.
2) HELEN [21]: There are 2000 facial images for training and 330 facial images for testing in the HELEN dataset, which are collected under wild conditions. We also download it from the IBUG website.
3) AFW [22]: There are 337 facial images in AFW dataset, which are collected from internet. We use all of them for training. And these images are downloaded from IBUG website in 68 landmarks annotation.

For all these datasets (total 3148 training images), we augmented the detected training images by random rotation (limited in $[-5^{\circ}\ 5^{\circ}]$), flipping and translation. Specifically, the bounding boxes we used are offered by IBUG website. We also used Normalized Root Mean Squared Error (NRMSE) as well as CFAN [14] to measure the error between the predicted landmark positions and the ground truth, which is normalized by distance between the eyes center in Euclidean metric. Finally, the Cumulative Error Distribution (CED) curves of NRMSE are used to evaluate the performances of different methods.

4.2. Investigation on 2-LCAN

Since our 2-LCAN has one 1st LCAN and three 2nd LCANs, we illustrate experimental result of each network separately and investigate the role of each network plays in the entire 2-LCAN. We conduct the experiments on the Labeled Face Parts in the Wild (LFPW) of 68 landmarks.

The experimental results are shown in Fig. 3. As seen, the CED of the 1st LCAN is 0.65 when NRMSE is 0.1. Obviously the shape estimation is not accurate enough, so we conduct the following refinements. In the 2nd LCAN 1, benefited from the 1st LCAN instead of using the mean shape as initialization, the CED is improved up to 23%. In the 2nd LCAN 2, the CED gains are 34% and 7% when NRMSEs are 0.05 and 0.1 respectively, which verify our assumption about the effects of smart training method. In the last 2nd LCAN, the improvements are slightly decreased because the smaller searching regions and more precise refinements with the higher and higher image resolutions.

4.3. Comparison with the State-of-the-Art Methods

We conduct our experiments on LFPW [20] and HELEN [21] dataset with 68 landmark points to compare our proposed methods with the state-of the-art methods, i.e., RCPR [9], SDM [11], CFAN [14] and DRMF [23]. For methods [9, 11, 23], we use the source codes that are published online. Since we cannot get available code of CFAN, we implement it by ourselves for comparing. The results are shown in Fig. 4 and Fig. 5. It can be seen that both of our method and CFAN perform better than other methods, which fully explains the advantages of deep learning methods on nonlinear tasks. For LFPW dataset, our method performs the best. When NRMSE is 0.05, our 2-LCAN outperforms CFAN by 11% improvements. Although the CED is similar to CFAN's when NRMSE is 0.1, there is only one test sample left when NRMSE is 0.15. For HELEN dataset, while the results are similar to CFAN's, our 2-LCAN still perform better, that is to say, our strategy of separating loss function and fusing SIFT and local pixels features together make our model more robust to the variations and occlusions.

5. CONCLUSION

In this paper, we have proposed a novel face alignment method based on 2-level cascaded auto-encoder networks. Our method can estimate an initial face shape according to the input and then refine the initial shape in a cascaded manner. Furthermore, we separate the loss functions of different facial components and fuse two features together to make our model more robust to the variations and occlusions. Finally, we also present a smart training method to prevent over-fitting in training process.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China (61402062) and Natural Science Foundation of Chongqing (cstc2015jcyjA1061).

6. REFERENCES

[1] C. Chen, A. Dantcheva, and A. Ross. Automatic facial makeup detection with application in face recognition. In *ICB*, pages 1–8, 2013. 1

[2] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful face–pain expression recognition using active appearance models. *IVC*, 27(12):1788-1796, 2009. 1

[3] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. In *IEEE Transactions on Patern Analysis ans Machine Intelligence*, 35(10):2401-2412, 2013.

[4] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. *Comput. Vis. Image Understand*. 61(1), 38 – 59, 1995.

[5] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *IEEE Trans.Pattern Anal. Mach. Intell*. 23(6), 681 – 685, 2001.

[6] Tzimiropoulos, G., Pantic, M.: Optimization problems for fast AAM fitting in the-wild. In *ICCV*, 2013.

[7] Zhou, F., Brandt, J., Lin, Z.: Exemplar-based graph matching for robust facial landmark localization. In *2013 IEEE Inter-national Conference on Computer Vision (ICCV)*, pp. 1025 – 1032. IEEE, 2013.

[8] Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1078 – 1085. IEEE, 2010.

[9] X. P. Burgos-Artizzu, P. Perona, and Dollár, P., "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis*., pp. 1513-1520, 2013.

[10] Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. *Int. J. Comput. Vis*. 107(2), 177 – 190, 2014.

[11] Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In *CVPR*, pp. 532 – 539. IEEE, 2013.

[12] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection". In *CVPR*, pp. 3476-3483, 2013.

[13] Lai H, Xiao S, Cui Z, et al. Deep Cascaded Regression for Face Alignment[J]. *Computer Science*, 2015.

[14] Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *ECCV*, pp. 1-16, 2014.

[15] Zhang, J., Kan, M., Shan, S., Chen, X.: Leveraging datasets with varying annotations for face alignment via deep regression network. In *ICCV*, 2015.

[16] Xiao S. Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks. In *ECCV*, pp. 57-72, 2016.

[17] Zhang J, Kan M, Shan S, et al. Occlusion-Free Face Align-ment: Deep Regression Networks Coupled with De-Corrupt Auto-Encoders. In *CVPR*, pp. 3428-3437, 2016.

[18] David G. Lowe, "Object Recognition from Local Scale-Invariant Features". In *ICCV*, pp. 1150, 1999.

[19] S. Ren, X. Cao, Y. Wei and J. Sun, "Face alignment at 3000 FPS via regression local binary features". In *CVPR*, pp. 1685-1692, 2014.

[20] Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[21] Le V., Brandt J., Lin Z., Bourdev L., Huang T.S.: Interactive Facial Feature Localization[C]. In *ECCV*, pp. 679-692, 2012.

[22] Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pp. 2879 – 2886. IEEE, 2012.

[23] Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In *CVPR*, pp. 3444–3451, 2013.