FACE ALIGNMENT BY DEEP CONVOLUTIONAL NETWORK WITH ADAPTIVE LEARNING RATE

Zhiwen Shao[1]*, Shouhong Ding*[1]*, Hengliang Zhu*[1]*, Chengjie Wang*[2]*, and Lizhuang Ma*[1]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]Tencent Incorporated, China
{shaozhiwen, feiben, hengliang_zhu}@sjtu.edu.cn, jasoncjwang@tencent.com, ma-lz@cs.sjtu.edu.cn

ABSTRACT

Deep convolutional network has been widely used in face recognition while not often used in face alignment. One of the most important reasons of this is the lack of training images annotated with landmarks due to fussy and time-consuming annotation work. To overcome this problem, we propose a novel data augmentation strategy. And we design an innovative training algorithm with adaptive learning rate for two iterative procedures, which helps the network to search an optimal solution. Our convolutional network can learn global high-level features and directly predict the coordinates of facial landmarks. Extensive evaluations show that our approach outperforms state-of-the-art methods especially in the condition of complex occlusion, pose, illumination and expression variations.

Index Terms— Deep convolutional network, data augmentation, adaptive learning rate

1. INTRODUCTION

Face alignment, namely detecting facial landmarks such as eyes and noses, is a preprocessing stage for tasks like face verification [1, 2], face recognition [3, 4] and face animation [5, 6]. It was extensively studied in these years [7, 8, 9, 10, 11] and achieved great success. However, when face images are taken with partial occlusion and large head pose variations, the localization of facial landmarks may become inaccurate. Deep convolutional network has been shown to be effective in extracting features and classification [1, 2, 12]. And it is proved to be robust to occlusions [13]. Therefore, we use deep convolutional network to directly predict the coordinates of facial landmarks.

Although deep convolutional network has a strong ability of learning feature, it needs to be trained from abundant samples. To make up for the lack of training images annotated with landmarks, we propose a novel data augmentation strategy including four operations: translation, rotation, horizontal flip and JPEG compression. In this way, the learned model will be robust to low quality and variations in pose rotation.

During training, the choice of learning rate is very important. We design an adaptive learning rate algorithm to train the network which learns a mapping from faces to coordinates of facial landmarks. Detailed experimental evaluations show that our approach outperforms state-of-the-art methods.

The remainder of this paper is organized as follows. In the next section, we discuss the related works of face alignment and analyse the characteristics of various approaches. In Section 3 , we introduce data augmentation briefly and illuminate the structure of our deep convolutional network, following which the training algorithm is elaborated. Several comparative experiments are carried out in Section 4 to show the precision and robustness of our model. Section 5 concludes this paper.

2. RELATED WORK

Significant progress on face alignment has been achieved in recent years. Conventional approaches can be divided into two categories: optimization-based and regression-based.

Optimization-based methods minimize the error between estimated face shape and the true face shape. It's very vital for the error function to be able to be optimized well. AAM [14, 15, 16] is a typical optimization-based method which reconstructs entire face using an appearance model and minimizes the texture residual to estimate the shape. It's well known that AAM is sensitive to the initialization of parameters. And the learned appearance models have limited capacity to adapt complex variations, so it may not generalize well on unseen faces.

Regression-based methods estimate landmark locations explicitly by learning a regression function with the input of image appearances. Xiong et al. [11] predict shape increment by applying linear regression on SIFT features. Both Cao et al. [7] and Burgos-Artizzu et al. [8] use boosted ferns to regress the shape increment with pixel-difference features. These methods mainly refine the prediction of the landmark locations iteratively from an initial estimate, thus the final re-

This work was supported by a joint project of Tencent BestImage and Shanghai Jiao Tong University. It was also sponsored by the National Natural Science Foundation of China (No. 61133009 and 61472245).

sult is highly dependent on the initialization. In contrast, our deep convolutional network takes raw face images as input without any initialization.

There are only a few methods based on deep learning so far. Sun et al. [9] estimate the positions of facial landmarks with three-level cascaded convolutional networks. Zhang et al. [17] train a deep convolutional network with multi-task learning to improve the generalization of face alignment. Our method requires neither cascaded networks nor multi-task learning, bringing about remarkable reduction in model complexity, whilst achieving better performance.

3. OUR APPROACH

In order to solve the problem of the lack of training images, we propose a novel data augmentation strategy. And we adaptively change learning rate with two iterative procedures to ensure the network converge well during training.

3.1. Data augmentation

Before starting the face alignment, we need to carry out face detection on the training images as preprocessing. Then we can acquire a face bounding box for each image. The data augmentation consists of three steps: translation and rotation; horizontal flip; JPEG compression, as shown in Figure 1.

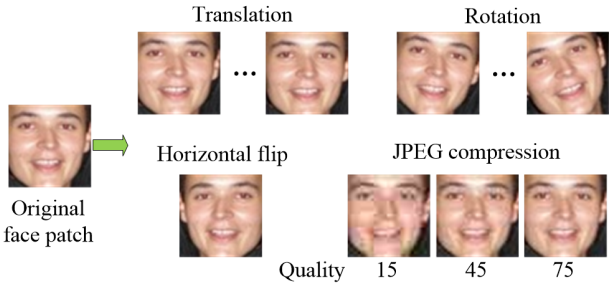

Fig. 1. New face patches derived from translation, rotation, horizontal flip and JPEG compression. The face patches are compressed with the JPEG qualities of 15, 45 and 75 respectively.

Compared with previous data methods, we combine a variety of augmentation strategies. During the first step, we slightly translate or rotate the face bounding box which is used for taking face patches. In this way, the training face patches are increased for dozens of times (34 in our experiments). It is worth mentioning that the new area contained in the face bounding box is derived from the original image rather than artificial setting, and the latter may has a bad impact on the training process of network. The translation operation helps to improve the robustness of landmark detection in the condition of tiny face shift, especially in face tracking. And our model can learn to adapt complex pose variation thanks to the rotation operation.

In the next steps, we horizontally flip each face patch and finally conduct JPEG compression with three different quality. Therefore, our network will be trained to be robust to poor-quality images which is ubiquitous in the real case.

3.2. Deep convolutional network

Our deep convolutional network contains eight convolutional layers followed by two fully-connected layers to learn global high-level features. And every two continuous convolutional layers connect with a max-pooling layer. The convolution operation is formulated as

$$y^j = \sum_i k^{ij} * x^i + b^j, \tag{1}$$

where x^i and y^j are the i-th input map and the j-th output map respectively. k^{ij} denotes the convolution kernel between the i-th input map and the j-th output map. b^j is the bias of the j-th output map. And $*$ denotes convolution. Max-pooling is expressed as

$$y^i_{j,k} = \max_{0 \le m,n < h} \{x^i_{j \cdot h+m, k \cdot h+n}\}, \tag{2}$$

where each $h \times h$ local region in the i-th input map x^i is pooled to be a neuron in the i-th output map. The network is based on VGG net [18] whose stacked multiple convolutional layers jointly form complex features. Figure 2 shows the detailed structure of our network.

In order to accelerate the training of network, we add a batch normalization layer [19] after each convolutional layer. Batch normalization is scaling and shifting the normalized input as

$$y = \gamma \hat{x} + \beta, \tag{3}$$

where $\hat{x} = \frac{x - E[x]}{\sqrt{Var[x]}}$, and the expectation and variance are computed over a mini-batch from the training dataset. After normalizing each convolutional layer, ReLU nonlinearity ($y = max(0, x)$)is added to speed up convergence. We don't operate ReLU on last two fully-connected layers in order to preserve important information. The network input is $50 \times 50 \times 3$ for color face patches. And the output of the last layer is predicted coordinates of five landmarks: left eye center (LE), right eye center (RE), nose tip(N), left mouth corner (LM) and right mouth corner (RM).

To guarantee numerical stability and reduce computational cost, we shrink the coordinates of landmarks with scale factor λ. Our network uses the Euclidean loss

$$L = \frac{1}{2}(f - \hat{f})^2, \tag{4}$$

where f is a vector that consists of ground truth, and $\hat{f}$ denotes predicted landmark locations. The gradient of loss L is

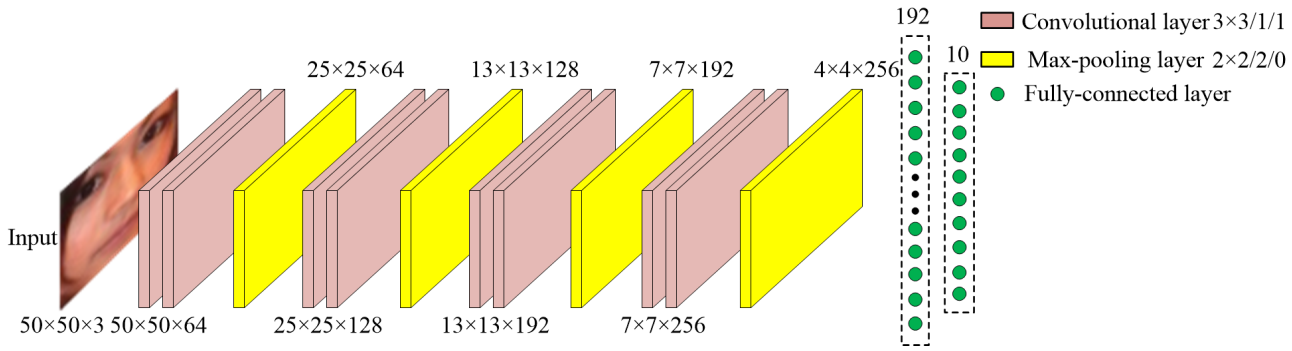

Fig. 2. The structure of our network. The equation $h \times w \times c$ beside each layer denotes that the dimension of map is $h \times w$ and the number of map is c. Every two continuous convolutional layers share the same equation. The equation $k_h \times k_w/s/p$ denotes that the filter size is $k_h \times k_w$, and the stride and padding of filter are s and p respectively. The filter parameters of each convolutional layer are identical, the same goes for max-pooling layers.

back-propagated to update network connection weights during training. And we need to magnify predicted landmark locations $\hat{f} = \hat{f}/\lambda$ as the final output.

3.3. Adaptive learning rate algorithm

Since the value of Euclidean loss may be several hundred or even several thousand, it is highly likely that computer numerical calculation scope may be exceeded in the process of back propagation. So the choice of learning rate is very important when training our network. We propose an innovative adaptive learning rate algorithm sketched in Algorithm 1.

Algorithm 1 The training algorithm with adaptive learning rate.

Input: Network N with trainable initialized parameters Θ_0, training set Ω, validation set Φ, control parameters α, t, k.

Output: Trainable parameters Θ.

1: Testing N and calculating the loss L_0 on Φ;
2: Setting the learning rate $\eta = \alpha/L_0$ and calculating the loss L on Ω;
3: **while** $L > t$ **do**
4: Training N with back propagation (BP) [20] algorithm and calculating L;
5: **if** l hasn't been reduced for k iterations **then**
6: $\eta = \eta \cdot 0.1$;
7: **end if**
8: **end while**
9: Setting $\eta = \alpha/L$;
10: **while** not convergence **do**
11: Executing step 4 to 7;
12: **end while**

During early training period, the network link weights will be changed sharply if learning rate is too large. Thus we firstly assign learning rate depended on initial testing loss. Then when network loss was reduced significantly, changing learning rate to be a larger value which will be decreased adaptively in subsequent training process.

It should be noted that Algorithm 1 consists of two iterative procedures for adaptive learning rate decrease. The training loss will be quickly reduced in the first procedure, leading to significant reduction in computational cost. During the second procedure, the network is convergent if loss tested on validation set is minimal and nearly unchanged.

4. EXPERIMENTS

We firstly investigate the advantages and effectiveness of our training algorithm by comparing to the algorithm with only one iterative procedure. Then we compare our approach with previous works on two public test sets, LFPW [21] and AFLW [22]. Our training and validation sets are identical to [9] and have no overlap with the test sets.

LFPW is collected from the web and contains $1,432$ face images which show large variations in pose, illumination, expression and occlusion. It shares only image URLs and some are no longer valid, so we use only $1,030$ images provided by [9].

AFLW includes $25,993$ face images gathered from Flickr. It is more challenging than other datasets such as LFPW. We use 2995 images of the dataset for testing as same as [17].

4.1. Algorithm discussions

Although the number of our original training images is only $10,000$, we can acquire dozens of times face patches, exactly totally $2800,000$ ($10,000 \times 35 \times 2 \times 4$) through translation and rotation, horizontal flip and JPEG compression in turn. When training our network, the control parameters α, t, k are set to be 0.015, 3 and $40,000$ respectively based on practical experience. If scale factor λ is too small, errors will be magnified excessively. So we assign it to be 0.2.

We train our network with adaptive learning rate algorithm and the algorithm with only one iterative procedure, respectively. When removing the second iterative procedure, the algorithm continuously runs the first iterative procedure until converging. The relationship between the loss tested on validation set and iterations of two different algorithms are shown in Figure 3.

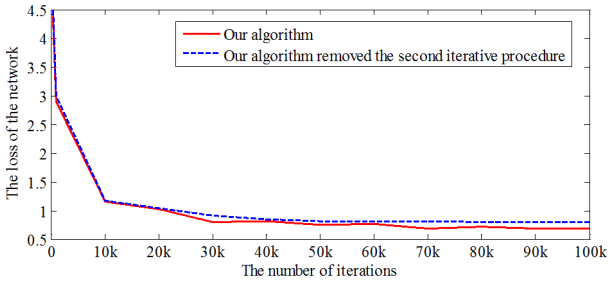

Fig. 3. The loss tested on validation set of our network vs. the number of iterations with different algorithms.

During early iterations, the loss with two algorithms are both decreased remarkably. But in later iterations, the loss of our algorithm is reduced to a smaller value. The minimal loss of our algorithm and the other are 0.6823 and 0.7938 respectively. Although the difference of loss is just 0.1115, the difference of average distance is approximately $\sqrt{0.1115/(\lambda^2) \times 2/5} \approx 1.0559$ in an image whose size is 50×50. Obviously the improvement is significant. Therefore our algorithm is highly likely to search a more optimal solution with a larger initial learning rate in the second procedure.

4.2. Comparison with other methods

We evaluate our approach based on mean error, similar to most previous works. The mean error is measured by the distances between estimated landmarks and the ground truths, normalized with the inter-pupil distance. We compare with state-of-the-art methods including ESR [7], RCPR [8], SDM [11], cascaded CNN [9] and TCDCN [17] as shown in Figure 4 and Table 1. The results of some methods are shown in original literatures or provided by later other literatures, and we implement other methods which didn't show their results on related datasets. It is obvious that our approach outperforms all the state-of-the-art methods and has a high accuracy for each landmark even on the challenge AFLW.

Compared with cascaded CNN and TCDCN, using deep convolutional network as same as ours, we require neither cascaded networks nor multi-task learning. Our method takes 67 ms to process an image on a single Intel Core i5 CPU, whilst cascaded CNN requires 120 ms. It's clear that our method is much faster.

Figure 5 shows several examples of landmark detection using cascaded CNN and our approach respectively. We observe that two methods both have a good performance on

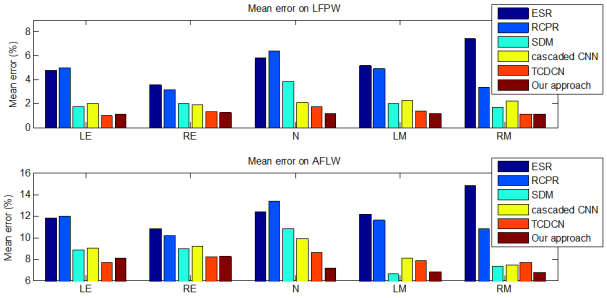

Fig. 4. Comparison of different methods on LFPW and AFLW: the mean error over each landmark.

Table 1. The mean error on LFPW and AFLW.

Method	LFPW	AFLW
ESR [7]	5.36	12.4
RCPR [8]	4.58	11.6
SDM [11]	2.26	8.5
cascaded CNN [9]	2.10	8.72
TCDCN [17]	1.33	8.0
Our approach	1.17	7.42

these challenge images, but ours achieves higher accuracy in detail. So the proposed method is robust to faces with complex variations in pose, illumination, expression, occlusion and quality.

Fig. 5. The results of cascaded CNN and our approach on several challenge images.

5. CONCLUSION

We propose an effective deep convolutional network based on data augmentation and adaptive learning rate for facial landmark detection. The former solves the lack of training images and the latter contributes to converging to a more optimal solution. Our approach directly predicts the coordinates of landmarks using single network without any other additional operations, whilst significantly improves the accuracy of face alignment. And we believe that the proposed data augmentation and training algorithm with adaptive learning rate can also be applied to other problems like face recognition.

6. REFERENCES

[1] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1701–1708.

[2] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[3] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning identity-preserving face space," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 113–120.

[4] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning multi-view representation for face recognition," *arXiv preprint arXiv:1406.6947*, 2014.

[5] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou, "3d shape regression for real-time facial animation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, pp. 41, 2013.

[6] Chen Cao, Qiming Hou, and Kun Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 43, 2014.

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[8] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár, "Robust face landmark estimation under occlusion," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1513–1520.

[9] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep convolutional network cascade for facial point detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3476–3483.

[10] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[11] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 532–539.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[13] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," *arXiv preprint arXiv:1412.1265*, 2014.

[14] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 6, pp. 681–685, 2001.

[15] Jason Saragih and Roland Goecke, "A nonlinear discriminative approach to aam fitting," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[16] Patrick Sauer, Timothy F Cootes, and Christopher J Taylor, "Accurate regression procedures for active appearance models.," in *BMVC*, 2011, pp. 1–11.

[17] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision–ECCV 2014*, pp. 94–108. Springer, 2014.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning internal representations by error propagation," Tech. Rep., DTIC Document, 1985.

[21] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Narendra Kumar, "Localizing parts of faces using a consensus of exemplars," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2930–2940, 2013.

[22] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.