Facial Landmark Localization by Part-Aware Deep Convolutional Network

Keke He and Xiangyang Xue

Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China
{kkhe15,xyxue}@fudan.edu.cn

Abstract. Facial landmark localization is a very challenging research task. The localization accuracy of landmarks on separate facial parts differ greatly due to texture and shape, however most existing methods fail to consider the part location of landmarks. To solve this problem, we propose a novel end-to-end regression framework using deep convolutional neural network (CNN). Our deep architecture first encodes the image into feature maps shared by all the landmarks. Then, these features are sent into two independent sub-network modules to regress contour landmarks and inner landmarks, respectively. Extensive evaluations conducted on 300-W benchmark dataset demonstrate the proposed deep framework achieves state-of-the-art results.

Keywords: Facial Landmark Localization, Deep Learning

1 Introduction

Facial landmark localization is to to automatically localize the facial key points including eyes, mouth, nose and other points on the face cheek. Due to its relevance to many facial analysis tasks like face recognition, face attribute analysis [10], 3D face modeling and etc, facial landmark localization has attracted increasing interests in the past years.

However, in an uncontrolled setting, face is likely to have large out-of-plane tilting, occlusion, illumination and expression variations. Facial landmark localization remains a challenging problem.

In general, existing methods to locate the facial landmarks can be divided into three categories: the first category is the ASM [6] and AAM [5] based methods, which fit a generative model by global facial appearance. However, these methods require expensive iterative steps and rely on good initializations. The mean shape is often used as the initialization, which may be far from the target position and hence inaccurate.

The second category is cascade regression based methods. The cascade regression framework has been proposed in recent works [17], which tries to estimate the facial landmark positions by a sequence of regression models. These methods obtain the coarse location first, and the following steps are to refine the initial estimate, yielding more accurate results. Sun [13] proposed three-level

cascaded deep convolutional networks. Zhou [19] designed a four-level coarse-to-fine network cascade to spread the network complexity and training burden of traditional convolutional networks. However these methods need to train individual systems for each group of the landmarks, the computational burden grows proportional to the group numbers and cascade levels. For example, this cascaded CNN method [13] needs to train 23 individual CNN networks.

Recently, a new framework based on multi-tasking has been proposed. The multi-task framework leverages other prior facial information like pose to assist landmark localization. Zhang [18] showed that learning the facial landmark localization task together with some correlated tasks, e.g., smile estimation would be beneficial. Yang [16] showed prior 3D head pose information will improve facial landmark localization accuracy. However these methods require auxiliary labels beyond landmarks and ignore the fact that landmarks on different facial parts have unbalanced localization difficulties. For instance, the detection of a mouth corner would be easy because there is abundant local information to capture. In contrast, the exact position of a landmark on the cheek is difficult to decide. As a consequence, it would be hard to optimize all the landmarks just in a single stage.

In this paper, we propose a novel end-to-end regression structure using deep convolutional neural network (CNN). Contrary to others, our method does not involve multiple individual models or require auxiliary labels. More importantly, the framework treats landmarks on different facial part differently which helps to learn discriminative features.

Our deep architecture first encodes the image into the feature maps shared by all the landmarks. Then, these shared features are sent into two individual sub-network modules to regress contour part landmarks and inner part landmarks respectively. The proposed method is called Part-Aware Convolutional Neural Network (PA-CNN).

To sum up, our main contributions are four-fold:

1) We propose a novel end-to-end regression CNN model for facial landmark localization by incorporating a contour landmark sub-network and an inner landmark sub-network into a unified architecture.

2) We clarify that all the landmarks sharing low level convolutional features and being independent in the latter layers can improve the accuracy and robustness.

3) We demonstrate what the sub-network learns by visualizing intermediate layer activations.

4) Finally, we show that the proposed network achieves state-of-the-art result on the 300-W [12] benchmark.

2 The Proposed Method

We will demonstrate our method in this section, describe the method in detail in 2.1 and analyze how to optimize our method in 2.2.

2.1 Architecture of PA-CNN

The proposed Part-Aware CNN (PA-CNN) framework integrates a contour landmark sub-network and an inner landmark sub-network, which handles the landmarks on different facial parts. Figure 1 illustrates the architecture of PA-CNN in detail. We divide the total 68 landmarks into 2 categories because of the difference in texture and shape. The inner landmark denotes the 51 landmarks for eyebrows, eyes, nose and mouth. The contour landmark is other 17 landmarks on the face contour.

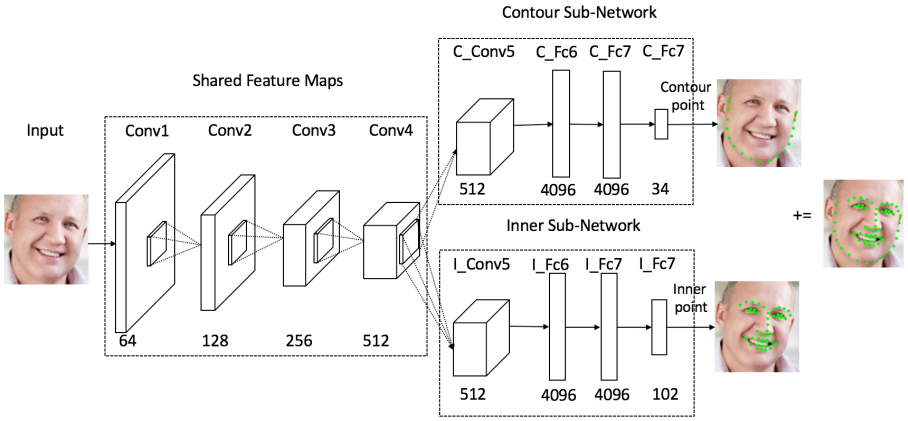

Fig. 1. PA-CNN network structure. This network structure uses cropped face detection color image as input. After four shared convolutional layers, the network branches into two sub-networks to localize contour landmarks and inner landmarks respectively. Those sub-networks contain one convolutional layer and three fully connected layers.

The PA-CNN uses cropped face color image as input, then following 4 convolutional layers and 2 pooling layers extract feature maps. In these stages, all the landmarks share the same weights. We have two concerns in designing bottom sharing convolutional layers. First, all the landmarks can incorporate general characteristics. By sharing input image and several convolutional layers, the context over the face can be utilized to locate each landmark. At the same time, all landmarks are implicitly encoded the geometric constraints. Second, sharing bottom layers makes our model time efficiency. In the early stage of the network, each layer extracts low level features. These features can be shared all across the face. If we use individual models, one for predicting 17 contour landmarks and another for predicting 51 inner landmarks, similar convolution modules need to densely convolve the entire image twice, it would be very time consuming.

After shared layers, the proposed network branches into two sub-networks. Each of the sub-networks takes the feature maps of the last shared convolutional layer as input, then through one convolutional layer and two 4096 dimension

fully connected layers. For the contour landmark sub-network, the dimension of the last fully connected layer is 34 because contour sub-network produces 17 landmarks and each landmark has x, y coordinates. The dimension of inner sub-network's last layer is calculated in the same way. The main concern of designing sub-networks is to impose our model to be specialized for specific landmarks. As inner landmarks and contour landmarks have unbalanced localization difficulties, learning to detect contour landmarks and inner landmarks respectively makes our model concentrate on the corresponding parts on the face. The integration of two sub-networks enables our PA-CNN to capture unique characteristics of landmarks. Later experiments in 3.3 will show this independence and share strategy is superior to optimal all landmarks in a single network stage. Finally, the outputs from the two sub-networks are combined to obtain the final result for each face.

2.2 Optimization

Let $x_i, y_i \in R$ be the x, y-coordinates of the i th facial landmark in an image I. Then the vector $[x_1, y_1, ..., x_N, y_N]^T$ denotes the coordinates of all the N facial landmarks in I, we take the vector P as the estimated landmarks and G as the ground truth landmarks. We define the landmark localization error as: $E = ||P - G||^2$. The predicted landmarks can be defined as

$$P = f(I; w) \tag{1}$$

f represents no-linear function, I donates the input image and w is the network weights. P can be calculated by network forward propagation. Finally, in a training batch, the network error can be represented as

$$E = \underset{w}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} ||f(I_i; w) - G_i||^2 \tag{2}$$

N donates the batch size, we set 70 in the training stage. The goal is to learn the optimal parameter w to minimize the loss function E by the training samples. According to the classical back propagate algorithm. For the exclusive layers like $conv5, fc1, fc2$, the weights are updated by the partial derivatives of loss with respect to weights $\Delta w = -r\frac{\partial E}{\partial w}$, r is the learning rate which controls the weight updating scale. For instance, the contour sub-network's $conv5$ layer can be updated by $\Delta w_{conv5} = -r\frac{\partial E_{con}}{\partial w_{conv5}}$, E_{con} is the loss of contour sub-network. The parameters in sub-network layers are only related to the sub-network loss. For the sharing layers like $conv1, ..., conv4$, the contour sub-network loss and the inner sub-network loss are back propagate together, can be represented as

$$\Delta w = -r\frac{\partial E_{con}}{\partial w} - \lambda * r\frac{\partial E_{inn}}{\partial w} \tag{3}$$

where E_{con} is the loss of contour sub-network, E_{inn} is the loss of inner sub-network, λ is the weight parameter to balance contour sub-network loss and inner sub-network loss, we choose 2 for our experiments.

3 Experiments

We conduct experiments on the 300-W dataset [12]. We will introduce this dataset in 3.1, describe the training detail in 3.2, and analyze the experiment results in 3.3, 3.4, 3.5, 3.6.

3.1 Datasets

300-W is short for 300 Faces in-the-Wild [12]. It's a commonly used benchmark for facial landmark localization problem. It is created from existing datasets, including LFPW [1], AFW [20], Helen [9] and a new dataset called IBUG. Each image has been densely annotated with 68 landmarks. Our training set consists of AFW, the training sets of LFPW and the training sets of Helen, with 3148 images in total. Our testing set consists of IBUG, the testing sets of LFPW and the testing set of Helen, with 689 images in total. IBUG subset is extremely challenging because its images have large variations in face pose, expression and illumination. The IBUG is called Challenging, testing sets of LFPW and the testing set of Helen are called Common, and all the 689 images are called Fullset.

Data augmentation. We train our models only using the data from the training data without external sources. We employ two distinct forms of data augmentation to enlarge the dataset. The first form of data augmentation is image rotations. We do random rotation on images with angles range in (-5°, 5°), (-10°, 10°), (-15°, 15°). The second form of data augmentation is random translation to right, left, up, down ranging in (-0.05, 0.05), which is a proportion of bounding box height or width. Finally, we enlarge the training data to 39561 images.

Evaluation. We evaluate the alignment accuracy by the popular mean error. The mean error is measured by the distances between the predicted landmarks and the ground truths, normalized by the inter-pupil distance, which can be represented as

$$err = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{M}\sum_{j=1}^{M}|p_{i,j}-g_{i,j}|_2}{|l_i-r_i|_2} \tag{4}$$

where N is the total test image number, M is the landmark number, p is the predicted landmark position, g is the ground truth landmark position. l_i, r_i are the left eye position and right eye position of i th image.

3.2 Implementation Detail

We implement the proposed method based on the open source Caffe [7] framework. We first crop the image using the bounding box with 0.05W padding on all sides (top, bottom, left, right), where W is the width of the bounding box. Then we resize the cropped image to 224×224.

We use the pre-trained VGG-S [4] model to initialize our network. The first four convolutional layers of the VGG-S network are used to initialize four shared

convolutional layers. These shared layers are designed to produce feature maps from the entire input image. The rest layers of the VGG-S network are used to initialize both the contour sub-network and the inner sub-network. Our framework is learnt in an end-to-end way. We set min-batch size 70, the weight decay parameter 0.0001 and training learning rate 0.0001. We train our models by stochastic gradient descent with 0.9 momentum. Training continues until converge.

3.3 Performance Analysis

We compare our PA-CNN with two distinct network structures. One outputs all the landmarks in the last layer, named as One-CNN. Another uses individual network for contour and inner landmarks, named as Two-CNN. Figure 2 shows the network detailed structures.

Accuracy. Table 1 shows the localization error of contour landmarks and inner landmarks. PA-CNN has performance improvements on both two kinds of landmarks comparing to One-CNN, which shows sub-network helps to learn discriminative features. PA-CNN is also comparable to Two-CNN, which demonstrates sharing several convolutional layers will not hurt accuracy but help to capture general features.

Efficiency. Table 2 lists the testing time of each method. By sharing several convolution layers, PA-CNN is faster than Two-CNN.

Figure 3 shows the error curve of PA-CNN. Our PA-CNN which concentrates on different parts of the facial landmarks and does not involve multiple individual networks is both accurate and efficient.

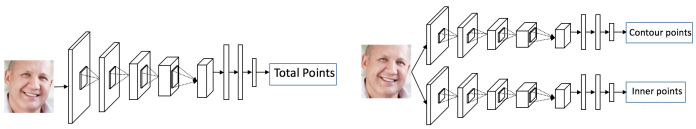

Fig. 2. Other two network structures. (a). Left is One-CNN structure. (b). Right is Two-CNN structure.

Table 1. Error of Landmarks on Fullset.

Methods	**Contour Error**($\times 10^{-2}$)	**Inner Error**($\times 10^{-2}$)
One-CNN	9.02	5.04
Two-CNN	8.95	4.81
PA-CNN	**8.85**	**4.79**

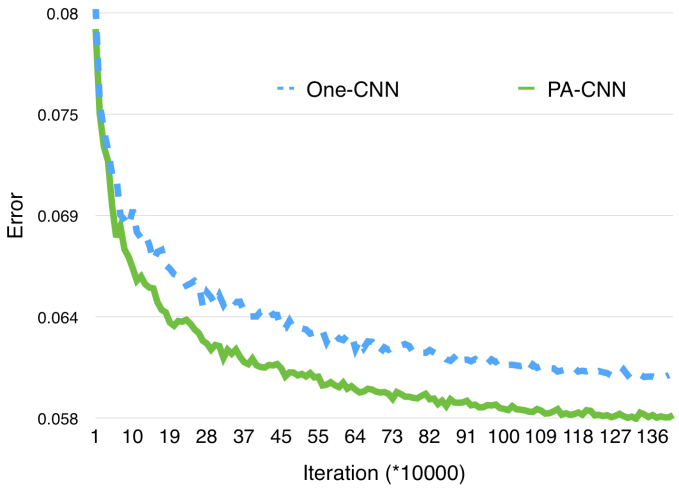

Fig. 3. Error Curve. Comparison of PA-CNN and One-CNN on FullSet test images. We can draw the conclusion that PA-CNN is superior to One-CNN on FullSet due to two sub-networks offering different concentrates on two kinds of facial landmarks.

Table 2. Comparison of testing time on Tesla K20 gpu.

Methods	**One-CNN**	**Two-CNN**	**PA-CNN**
Time(s/image)	0.0180	0.0359	0.0300

3.4 Intermediate Feature Visualization

We gain intuition about what our network learns by visualization intermediate features. We plot the feature map of *conv*5 layer at figure 4. Contour sub-network and inner sub-network have own *conv*5 layer. The first column is the input image. The second column is the activation of contour sub-network. The third column is the activation of inner sub-network. We can see that in contour sub-network's feature map, the high activations are always near the image contour, which indicates that contour sub-network concentrates on contour parts. In the inner sub-network, the high activation parts are corresponding to the eyes and nose parts, which reveal inner sub-network pays more attention on the inner facial parts. This result demonstrates that two sub-networks explore the different characteristics of landmarks on various parts dramatically and help to learn the discriminative features.

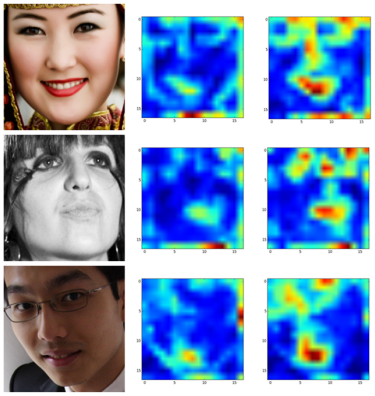

Fig. 4. The activation of sub-network. The first column is the input image. The second column is the activation of contour sub-network *conv*5 layer, high activation parts are around corner. The third column is the activation of inner sub-network *conv*5 layer, high activations are corresponding to the eyes and nose parts. (Best view in electronic form.)

3.5 Comparison with State-of-the-arts

In all previous analysis, we use VGG-S to pre-train our framework, and draw a conclusion that PA-CNN which shares the low-level convolutional features and uses two sub-network to localize different type of landmarks is superior to network in figure 2. We also compare the result of PA-CNN with the existing public methods, including PCRR [2], GN-DPM [14], CFAN[17], ESR [3], SDM [15], ERT [8], LBF [11]. The overall experimental results are reported in Table 3. We improve the performance on the 300-W dataset, especially on Challenging test set.

Table 3. The evaluation on 300-W dataset

Methods	**Common**($\times 10^{-2}$)	**Challenging**($\times 10^{-2}$)	**Fullset**($\times 10^{-2}$)
PCRR [2]	6.18	17.26	8.35
GN-DPM [14]	5.78	-	-
CFAN [17]	5.50	16.78	7.69
ESR [3]	5.28	17.00	7.58
SDM [15]	5.57	15.4	7.5
ERT [8]	-	-	6.4
LBF [11]	4.95	11.98	6.32
PA-CNN	**4.82**	**9.80**	**5.79**

3.6 Localization Results

Figure 5 shows the result of our localization method on the 300-W test images. Even the testing images have large head poses or occlusions, our method is accurate and robust.

Fig. 5. Facial landmark localization result on test images.

4 Conclusions

In this paper, we propose a novel end-to-end regression structure using convolutional neural network to deal with different facial landmarks. Our deep architecture first encodes the image into feature maps shared by all the landmarks. Then, these features are sent into two individual sub-network modules to regress contour landmarks and inner landmarks respectively. Experimental results on challenging 300-W dataset demonstrate our approach achieves state-of-the-art result. In future, we will extend the proposed PA-CNN to more facial parts.

Acknowledgments. This work was supported in part by two STCSM's Programs. (No. 15511104402 15JC1400103)

References

1. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(12), 2930–2940 (2013)
2. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 1513–1520. IEEE (2013)

3. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision 107(2), 177–190 (2014)
4. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Computer VisionECCV98, pp. 484–498. Springer (1998)
6. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding 61(1), 38–59 (1995)
7. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
8. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1867–1874. IEEE (2014)
9. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Computer Vision–ECCV 2012, pp. 679–692. Springer (2012)
10. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3730–3738 (2015)
11. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1685–1692. IEEE (2014)
12. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. pp. 397–403. IEEE (2013)
13. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 3476–3483. IEEE (2013)
14. Tzimiropoulos, G., Pantic, M.: Gauss-newton deformable part models for face alignment in-the-wild. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1851–1858. IEEE (2014)
15. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 532–539. IEEE (2013)
16. Yang, H., Mou, W., Zhang, Y., Patras, I., Gunes, H., Robinson, P.: Face alignment assisted by head pose estimation. arXiv preprint arXiv:1507.03148 (2015)
17. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: Computer Vision–ECCV 2014, pp. 1–16. Springer (2014)
18. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Computer Vision–ECCV 2014, pp. 94–108. Springer (2014)
19. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. pp. 386–391. IEEE (2013)
20. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2879–2886. IEEE (2012)