A DEEP FACIAL LANDMARKS DETECTION WITH FACIAL CONTOUR AND FACIAL COMPONENTS CONSTRAINT

Wissam J. Baddar, Jisoo Son, Dae Hoe Kim, Seong Tae Kim, and Yong Man Ro

Image and Video Systems Lab., School of Electrical Engineering, KAIST, Daejeon, 305-701, Republic of Korea

ABSTRACT

In this paper, we propose a new facial landmarks detection method based on deep learning with facial contour and facial components constraints. The proposed deep convolutional neural networks (DCNNs) for facial landmark detection consists of two deep networks: one DCNN is to detect landmarks constrained on the facial contour and the other is to detect landmarks constrained on facial components. A novel DCNN structure for the landmarks detection with facial component constraints is proposed, which branches the network at higher layers in order to capture the intricate local facial components features. Moreover, a novel learning strategy is proposed to learn the DCNN for detecting the landmarks on the facial contour by exploiting the relationship between facial contour landmarks and those on facial components. Experimental results have shown that the proposed method outperforms the state-of-the-art FLD methods.

Index Terms— Facial Landmark Detection, Deep Learning, Convolutional Neural Network

1. INTRODUCTION

Facial landmarks located around facial contour and facial components deliver semantic meanings of those components and the subjects face [1]. As a vital step for diverse face related applications, facial landmark detection (FLD) has received much attention in the computer vision area. Nonetheless, FLD is still a challenging problem in real world applications with large variations in face appearance, head pose, expressions, illumination, and partial occlusions [1].

FLD methods can be divided into two main approaches: optimization-based methods [2-6] and regression-based methods [7-12]. Optimization-based methods build a generative model during training. At the test stage, model parameters are iteratively updated to fit the model to the test image. The optimization-based methods try to tackle partial occlusions of landmarks by applying shape constraints. However, the shape constraints are relatively weak and cannot fully model the real-world complex variations of face images [13]. On the other hand, regression-based methods deal with FLD as a regression problem, which directly maps a test image to the coordinates of landmarks with a learned model. This avoids iterative optimization during testing. Regression-based methods usually build a mean shape of training set to apply as an initial value, which can result less-accurate detection when the shape of the test image is far from the mean shape [14].

With recent success of deep learning in a variety of computer vision problems, deep learning regression-based FLD methods have been proposed. Sun [15] proposed a cascaded deep convolutional neural network (DCNN) of three-stages. However, facial landmarks were independently refined after the first stage, which could distort the whole shape [16]. Zhang [16] proposed a coarse-to-fine auto-encoder network (CFAN), which consisted of stacked auto-encoders that initialized landmarks locations in a low resolution face image, and refined the landmark positions on higher resolution face images. CFAN globally detects landmarks, which could not fully consider variations of local facial components. Zhang [17, 18] applied multi-task learning to improve FLD performance, by learning a set of related auxiliary tasks that can aid FLD. Although multitasking improved the FLD performance, it requires acquiring additional auxiliary labels beforehand to train the FLD task.

In this paper, a new FLD method is proposed based on DCNN with facial contour and facial components constraints. Landmarks on the facial contour are significantly difficult to detect [19]. Facial components constraints are considered to learn the DCNN for FLD. The proposed deep network consists of two DCNNs which are trained separately: 1) DCNN-C, trained to detect landmarks on the facial contour, and 2) DCNN-I, trained to detect landmarks on facial components. By separately learning DCNN-C and DCNN-I we improve the detection of landmarks on facial components. In addition, by jointly learning the landmarks on facial contour with the facial components, we improve the landmarks detection of facial contour. A novel learning strategy for DCNN-C is devised to jointly learn the DCNN-C hyper parameters, and transfer them to a DCNN dedicated only for landmarks detection on facial contour. In DCNN-I, a novel CNN structure is proposed to encode the relationships between landmarks on facial components. To that end, DCNN-I consists of shared lower layers and forks into branches at higher layers dedicated for landmarks on different facial components (eyes, eyebrows, nose, nose-bridge and mouth). The hyper parameters (filters) of the separate higher layers are learned only by the corresponding facial component, making the network more robust to partial occlusion, expression variations or head pose variations.

Moreover, sharing the lower layers maintains the detection of the overall shape in facial landmarks. The effectiveness of the proposed method is validated via comparative experiments on the publicly available 300-W dataset [20]. The experimental results show that the proposed method outperforms the state-of-the-art methods.

The reminder of this paper is organized as follows. In section 2, the proposed FLD with facial contour and facial components constraints is described. Section 3 discusses the experimental design and the results. Finally conclusions are drawn in section 4.

2. PROPOSED FACIAL LANDMARKS DETECTION

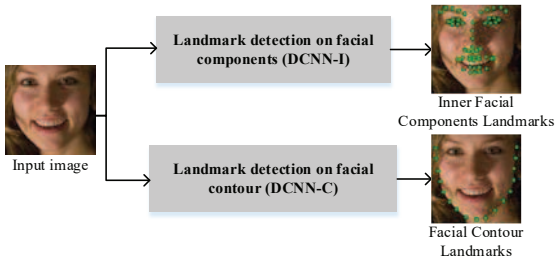

Fig.1: An overview of the proposed FLD with facial contour and facial components constraint.

Figure 1 shows an overview of the proposed FLD method with facial contour and facial components constraints. The proposed method is composed of two separate DCNNs, i.e., DCNN-I and DCNN-C, dedicated for detecting landmarks on facial components and those on facial contour, respectively. This separate design considers the facial contour and facial components properties, and applies components constraints that lead to improved FLD. The details of DCNN-I and DCNN-C designs, and their learning strategies are described in the following subsections.

2.1 Landmarks Detection on Facial Components in DCNN-I

The landmarks on facial components can be detected as a linear regression problem, where a loss function is the total loss from all landmarks in a batch of N images. The loss function can be written as

$$E = \frac{1}{2}\sum_{i=1}^{N}\left\|\mathbf{y}_i - f(\mathbf{x}_i;\mathbf{W})\right\|^2 , \qquad (1)$$

where $\mathbf{y}_i$ is a vector of all landmarks coordinates on facial components, $\mathbf{x}_i$ represents an input image and $f(.)$ is a function of $\mathbf{x}_i$ parameterized by the learned hyper parameters $\mathbf{W}$. This regression formulation assumes that the loss from all the landmarks collectively contributes to learn the hyper parameters of the DCNN. It also assumes that the relationships between all landmarks on facial components are considered jointly. However, the relationships between landmarks on facial components can be characterized by inter-facial component relationship and intra-facial component relationship. In the inter-facial component relationship, landmarks on each facial component are constrained by the location of the corresponding facial component, e.g. with the known location of nose landmarks, we can roughly estimate the location of eyes landmarks, etc. In the intra-facial component relationship, facial component variations can affect its own local landmarks but do not directly affect other facial component landmarks, e.g. when blinking, eye landmarks are affected but not the mouth landmarks.

To effectively incorporate the inter- and intra- facial component relationship between landmarks, we propose a novel DCNN structure that is composed of shared lower layers that fork into branches at the higher layers. The shared lower layers jointly learn the landmarks on facial components, incorporating the inter-facial component relationship between all landmarks on facial components. The higher branches are separated based on the intra-facial component relationship (within each facial component of eyebrows, eyes, nose-bridge, nose or mouth). This way, the higher layer hyper parameters are fine-tuned independently from other facial components landmarks to capture the intricate local features of the facial component. Moreover, this separation in learning the features results in an improved FLD that is more robust to occlusion and local variations.

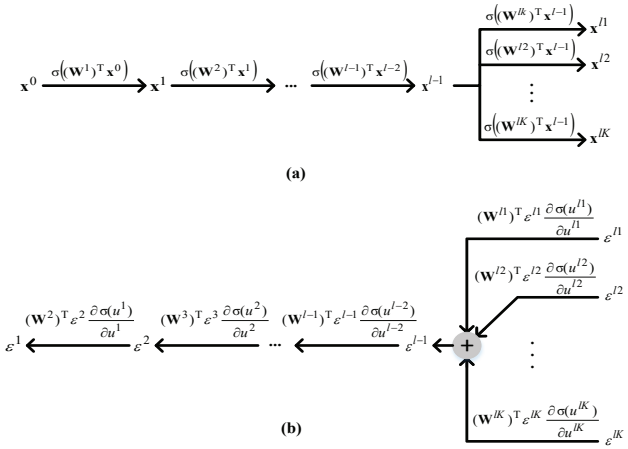

Fig.2: The learning process of DCNN-I constrained on facial components (a) DCNN-I Forward propagation (b) DCNN-I backward propagation and filter updates.

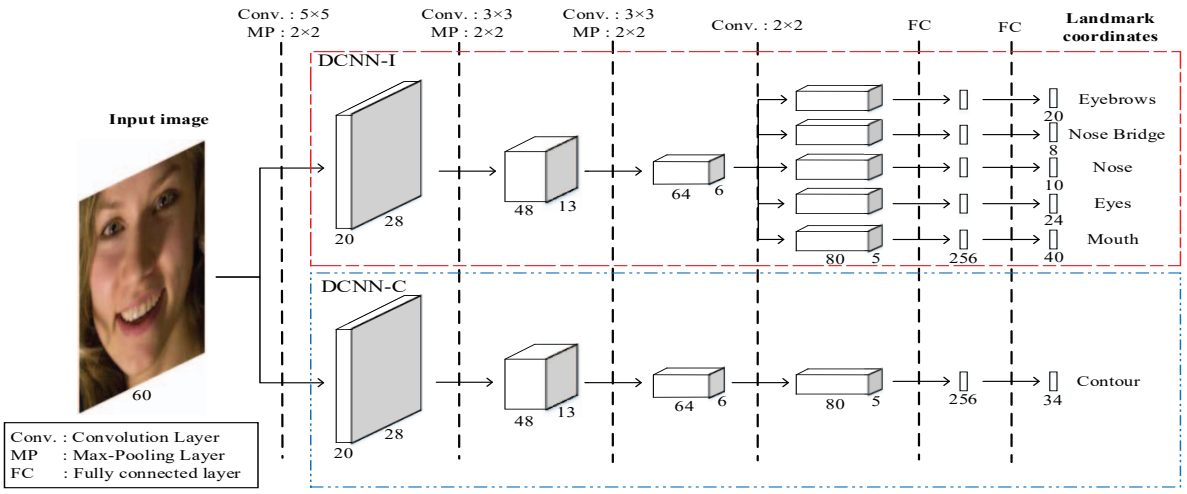

Fig.3: Details of the implemented network structure of the proposed FLD with facial contour and facial components constraints.

Figure 2 details the learning procedure of the proposed structure by describing the forward and the backward propagation passes. In a forward propagation pass (Fig.2 (a)), the shared lower layers propagate features in similar manner to a conventional DCNN, i.e., the features $\mathbf{x}^l$ at layer l are obtained by applying an activation function $\sigma(.)$ parameterized by $\mathbf{W}^l$ to the previous layer features $\mathbf{x}^{l-1}$. In the *K* branches of the higher layers, corresponding to *K* local facial components, each layer takes a replica of the feature map at the last shared lower layer, and propagates it according to each facial component. Note that $\mathbf{x}^{lK}$ in Fig. 2 (a) corresponds the latent features extracted from the *K*-th local facial component in the *l*-th layer.

To update the filters in a backward propagation pass (Fig.2 (b)), the higher layers loss from the *k*-th branch is independently back propagated by

$$\varepsilon^{lk} = \frac{\partial E^k}{\partial \mathbf{W}^k} = (\mathbf{y}_i - (\mathbf{W}^{lk})^{\mathrm{T}} \mathbf{x}_i)\mathbf{x}_i^{\mathrm{T}}, \qquad (2)$$

which update the filters corresponding to each facial component. The errors of the layers bellow the loss layer are computed by $\varepsilon^{l-1} = (\mathbf{W}^l)^{\mathrm{T}} \varepsilon^l \frac{\partial \sigma(u^l)}{\partial u^l}$ where $\frac{\partial \sigma(u^l)}{\partial u^l}$ is the gradient of the *l*-th layer activation function. Please note that at the point of branching (the last shared lower layer), the errors of all the local facial component branches are integrated and jointly back propagated, adjusting the lower layers filters based on the inter-facial component relationship between landmarks.

2.2 Landmarks Detection on Facial Contour in DCNN-C

Facial contour landmarks are significantly difficult to be detected due to facial background noise, head pose variations or even subject appearance variations such as facial hair [19]. To improve facial contour landmarks detection, learning them jointly with facial components can improve their detection, because facial components can depict the head pose and provide a rough estimate of the location of facial contour [21]. To improve the detection of facial contour landmarks without affecting the landmarks on facial components, we utilize a separate DCNN for detecting the facial contour landmarks (DCNN-C). In addition, we propose pre-training DCNN-C jointly with the facial components landmarks as follows.

For pre-training DCNN-C, a network similar in structure to DCNN-I is utilized and trained with all facial landmarks including facial contour landmarks with the same training process described in section 2.1. After the learning process is complete, the learned hyper parameters for the shared lower layers can be directly transferred to the lower layers of DCNN-C. For the higher convolutional layers (in this paper, the last convolutional layer and the fully connected layers) only the hyper parameters of facial contour landmarks branch are transferred to DCNN-C, which reduces the computation time at the test stage.

3. EXPERIMENS

3.1 Experimental Setup

To evaluate the proposed method, experiments have been conducted on the benchmark dataset, 300-W dataset [20]. The 300-W dataset is composed of a combination of existing datasets; LFPW [22] AFW [23] HELEN [24] and an additional subset of IBUG [20] which contains challenging head poses, expressions, illuminations and occlusions. The 68 landmarks provided for each face in the dataset were used as ground truths. The conducted experiments followed the same protocol in [18]. The data set was divided into 3,148 training images (collected from AFW, and the training sets of LFPW and HELEN) and test images consisting of 689 face images (collected from the test sets of LFPW, HELEN and challenging images collected from IBUG).

To successfully train the proposed DCNN models and to avoid over-fitting, a large number of training samples is required. Therefore, data augmentation was performed to increase the size of the training dataset from 3,148 face images to 40,082 face images by performing random

translation with ±3 pixels in both x and y directions, in-plane rotation in the range of ±50° with 2º step size, and zoomed by x1.1 to 1.2 zooming factors with a step size of x0.002.

To implement the proposed DCNN structure, Caffe [25] was utilized. The implemented network structure is shown in Fig.3. As shown, DCNN-I was composed of three shared lower layers, each of which consisted of a convolutional layer followed by a max pooling layer. The network forked into five local facial component branches at the fourth layer. Each branch consisted of a convolutional layer followed by two fully connected layers. The rectified linear unit (ReLU) was used as an activation function for each layer. DCNN-C, consisted of four convolutional layers, the first three of which were followed by max-pooling layers, and the fourth layer was followed by two fully-connected layers.

To evaluate the accuracy of the detected landmarks, mean error was used, which measures the distance between the estimated landmarks and the ground truths normalized by the inter-ocular distance. Eq. (3) shows the definition of the mean error, where M denotes the number of landmarks, $\mathbf{p}$ is the landmark prediction, $\mathbf{g}$ is the ground truth and l and r are the positions of the left eye and right eyes, respectively.

$$error = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{M}\sum_{j=1}^{M}\left|\mathrm{p}_{i,j}-\mathrm{g}_{i,j}\right|_2}{\left|l_i-r_i\right|_2}, \qquad (3)$$

3.2 Experimental Results

The first experiment verified the effectiveness of utilizing the facial contour and facial components constraints to improve the FLD performance. For comparison, a conventional DCNN was constructed with a four convolutional layer architecture similar to DCNN-C. Fig.4 shows the mean error comparison between the proposed DCNN and the conventional DCNN. The figure shows that the proposed method outperforms the conventional DCNN in detecting all facial components landmarks. For the full set of 68 facial landmarks, the proposed method achieves 6.12 mean error compared to 6.99 in conventional DCNN (Table 1). These results are in line with the fact that separating higher layers during the training of the DCNN utilizes the facial contour and facial components constraints, resulting the improved FLD that is robust to partial occlusion, expression variations or head pose variations.

In the second experiment, the FLD accuracy of the proposed method was compared to state-of-the-art FLD methods. Results from both hand-crafted FLD methods and deep learning FLD methods are presented in Table 1. As shown in the table, the proposed method outperforms both hand-crafted and deep learning state-of-the-art methods. These results indicate that the proposed method encodes the inter-facial component relationship between facial landmarks, and further improves the performance by utilizing the intra-facial component relationship.

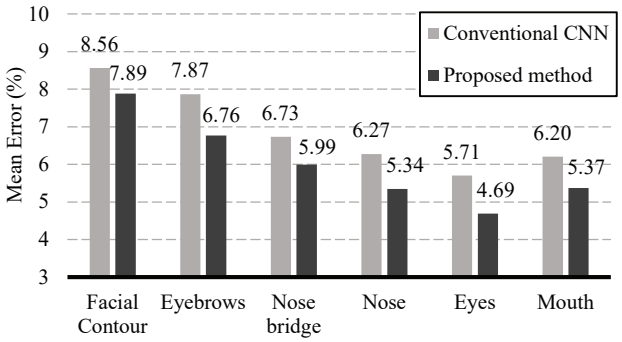

Fig. 4: FLD evaluation. Mean error comparative results between the proposed method and a conventional DCNN.

Table 1. Comparison with state-of-the-art methods

	Method	Mean error (%)
Hand-crafted	ESR[7]	7.58
	SDM [12]	7.50
	ERT [10]	6.40
	LBF [8]	6.32
Deep learning	CFAN [16]	7.69
	TCDCN [17]	6.83
	TCDCN-Averaged [26]	6.29
	Proposed Method	**6.12**

4. CONCLUSIONS

In this paper, a FLD method that utilizes facial contour and facial components constraints was introduced. Two separate DCNNs (DCNN-I and DCNN-C) were designed to detect landmarks on facial components and facial contour landmarks. For facial components landmarks detection, DCNN-I was designed with a novel structure that branched at higher level layers. This design was able to capture facial component variations, and fine-tune the corresponding hyper parameters independently, while the shared layers maintained the inter-facial component structure of facial landmarks. DCNN-C followed a conventional DCNN design, yet it was pre-trained jointly with facial components to improve the detection of the facial contour landmarks. The effectiveness of the proposed method was validated on the standard 300-W dataset. Experimental results showed that the proposed method outperforms state-of-the-art methods.

4. ACKNOWLEDGEMENT

This work was partially supported by the ICT R&D program of MSIP/IITP, grant No. B0101-16-0525.

REFERENCES

[1] N. Wang, X. Gao, D. Tao, and X. Li, "Facial Feature Point Detection: A Comprehensive Survey," *arXiv preprint arXiv:1410.1037,* 2014.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding,* vol. 61, pp. 38-59, 1995.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* pp. 681-685, 2001.

[4] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision,* vol. 60, pp. 135-164, 2004.

[5] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *Computer Vision–ECCV 2008*, ed: Springer, 2008, pp. 72-85.

[6] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition,* vol. 41, pp. 3054-3067, 2008.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision,* vol. 107, pp. 177-190, 2014.

[8] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 1685-1692.

[9] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, 2013, pp. 392-396.

[10] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 1867-1874.

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

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

[13] S. Zhang, H. Yang, and Z. Yin, "Multiple deep convolutional neural networks averaging for face alignment," *Journal of Electronic Imaging,* vol. 24, pp. 033013-033013, 2015.

[14] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson, "Face Alignment Assisted by Head Pose Estimation," *arXiv preprint arXiv:1507.03148,* 2015.

[15] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3476-3483.

[16] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *Computer Vision–ECCV 2014*, ed: Springer, 2014, pp. 1-16.

[17] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision–ECCV 2014*, ed: Springer, 2014, pp. 94-108.

[18] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," 2015.

[19] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, 2013, pp. 386-391.

[20] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, 2013, pp. 397-403.

[21] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 607-626, 2009.

[22] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 35, pp. 2930-2940, 2013.

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

[24] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Computer Vision–ECCV 2012*, ed: Springer, 2012, pp. 679-692.

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick*, et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 675-678.

[26] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning and transferring multi-task deep representation for face alignment," *arXiv preprint arXiv:1408.3967,* 2014.