LEARNING DEFORMABLE HOURGLASS NETWORKS (DHGN) FOR UNCONSTRAINED FACE ALIGNMENT

Jiaqiang Zhang, Congcong Zhu*, Suping Wu, Zhenhua Yu, Xuehong Sun, Hao Liu*

School of Information Engineering, Ningxia University, Yinchuan, 750021, China

ABSTRACT

In this paper, we propose a deformable hourglass networks (DHGN) approach to investigate the problem of face alignment, especially in such challenging cases when faces undergo large variations including severe poses, diverse expressions and partial occlusions in unconstrained environments. Unlike conventional feature extractions which cannot explicitly exploit irregular geometric structures for facial shapes, our DHGN learns a deformable mask to reduce the variances of facial deformation and extract attentional facial regions for robust feature representation. To achieve this, we carefully design a differential module, dubbed the deformable transformer, which typically incorporates with a regression subnet to predict a set of offsets and a masking operator to filter the semantic facial parts for feature representation learning. To further reinforce the alignment performance, we integrate our designed modules in the paradigm of stacked hourglass networks and jointly optimize the network parameters in an end-to-end manner. Extensive experimental results demonstrate very compelling performance in comparisons to most state-of-the-art methods.

Index Terms— Face alignment, deep learning, hourglass network, spatial transformer, biometrics.

1. INTRODUCTION

The basic goal of face alignment (*a.k.a.*, facial landmark localization) aims to detect a set of facial landmarks based on the facial appearance of the input image, which plays a significant pre-processing step for many facial attribute analysis tasks such as head pose estimation [1], facial expression [2], 3D geometric reconstruction [3], visual tracking [4, 5], etc. While many efforts have been devoted in recent literatures [6–12], the performance still remains unsatisfactory in practice especially when faces were captured in unconstrained environments. This is mainly because the relationship between face appearances and the variations of facial shapes is complexly nonlinear due to many difficulties, *e.g.*, large poses, varying expressions and severe occlusions. This quite motivates us to propose a robust face alignment approach versus sets of unconstrained scenarios.

Fig. 1. Our proposed DHGN versus conventional masks.

Cascade regression has emerged a dominant role in the areas of face alignment recently [6–9], which targets on seeking a series of discriminative mappings between local features and facial shape coordinates. By doing this, the facial shape is refined starting from the initialization in a coarse-to-fine manner. For examples, Xiong *et al.* proposed a supervised descent method (SDM) [7] which learns a set of linear mappings from the extracted local SIFT features by using the gradient descent method. To further improve SDM, Trigeorgis *et al.* developed a Mnemonic Descent Method (MDM) [13], making a cascade of networks to exploit the nonlinear relationship between face appearance and shape variations, which achieves promising performance by a large margin. In the term of the feature extraction method, they usually utilize the shape-index mask [9,14] to sample face regions with fixed geometric structures and then extract local features on them for coordinate regression. Moreover, these methods constrainedly explore partial features and even ignore the shape-sensitive information, which likely falls into local optima when the initial shape is far away from the target shape. To circumvent

*indicates equal contribution, the corresponding author is Hao Liu (e-mail: liuhao@nxu.edu.cn). This work was supported in part by the National Science Foundation of China under Grant 61662059 and Grant 61806104, in part by the Natural Science Foundation of Ningxia under Grant 2018AAC03035, in part by the Scientific Research Projects of Colleges and Universities of Ningxia under Grant NGY2018050, and in part by the Youth Science and Technology Talents Enrollment Projects of Ningxia under Grant TJGC2018028.

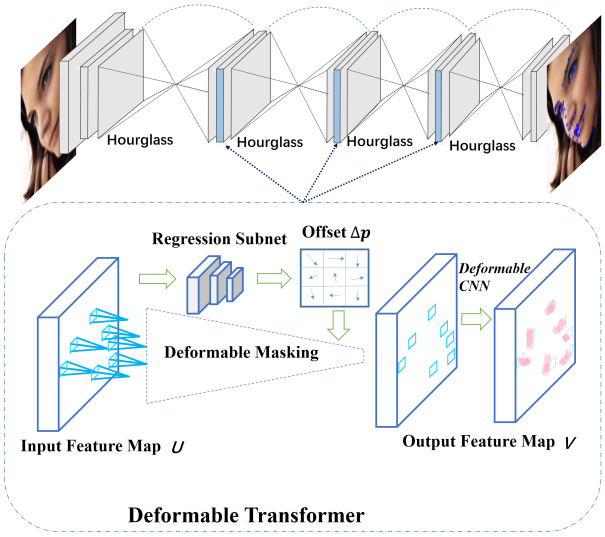

Fig. 2. Architecture of our DHGN.

this problem, heat map-based regression models [10, 11] have been proposed to take in the whole global face image and deploy an symmetric encoder-decoder deep architecture for facial landmark localization. Another major benefit of these methods is that they regard the heat maps as the regression targets, which significantly preserves the spatial constraints and relatively improves the performance [15, 16]. However, methods in this category apply convolutional operation with a fixed geometric structures, which likely losses shape-informative details and cannot explicitly exploit the facial shape variations in such cases when faces were captured in wild conditions. Fig. 1 visualizes some failure cases where the features are extracted via the limited fixed geometric structures. Taken two face alignment methods CFAN [9] and LBF [14] as the compared baselines, where CFAN provides a set of rectangles to extract shape-sensitive patches to exploit local features, and LBF learns the random pixel difference as the local features to preserve the shape-index constraint. Both methods obviously loss shape-informative details during feature extracting, which may give rise to the bias prediction. Quite differently from both methods, our DHGN automatically explores deformable ROI regions to exploit discriminative features for robust face alignment. It can be seen that the resulting alignment samples obviously highlight the superiority of our proposed approach.

In our approach, we propose a deformable hourglass networks (DHGN) approach for robust face alignment especially when faces were captured in unconstrained conditions. Different from existing face alignment approaches which integrates with fixed geometric structures for feature extraction, our DHGN develops an enhanced deformable transformer module to explore the attentional facial parts with irregular receptive fields. Specifically, our designed transformer incorporates with two key modules including the transformation regression subnet and a masking operator. The transformation regression subnet is designed to estimate a set of offsets, which aims to refine the positions of observation windows. Accordingly, the masking operator automatically crops the attentional regions based on the observation windows and provides plausible shape-sensitive information for robust feature representations. It should be noted that the transformer module is totally differential throughout the whole deep architecture. Having obtained the extracted attentional parts, we apply deformable convolution neural networks on them to reinforce the deformation of the receptive fields. Finally, we integrate the hourglass networks with our transformer module and stacked them to further improve the alignment performance. During training process, the network parameters are achieved by using the standard back-propagation algorithm in an end-to-end manner. Fig. 2 illustrates the detailed flowchart of our DHGN. To evaluate the effectiveness of the proposed approach, we conduct sets of experiments on four benchmark datasets and the results indicate very competitive performance compared with other state-of-the-art methods.

The main contributions of this work are summarized in the following two-fold aspects:

1) Compared with conventional regression-based face alignment methods, we carefully design a deformable transformer module to learn a set of offsets by directly feeding the input of feature maps. With these offsets, the deformable masking is leveraged to exploit the attentional face regions and achieves discriminative shape-informative features for robust face alignment.

2) The developed transformer module is approximately differential and can be readily integrated in the backbone alignment networks (*e.g.*, hourglass). Hence, this end-to-end training schema teaches an optimal optimization target and enhances the capacity of the regression model, which further reinforces the alignment performance.

2. DEFORMABLE HOURGLASS NETWORKS

Unlike existing deep learning-based methods that are inherently limited to model the deformable geometric structures, the basic idea of our proposed *Deformable Hourglass Networks* (DHGN) is to enhance the spatial sampling locations in the modules with additional offsets and learning these offsets from the alignment error function, without using any auxiliary supervision signals. The proposed module can readily integrate with the existing hourglass networks [16] architecture which was designed adapted for pose estimation. Hence, the whole networks can be easily trained end-to-end by standard back-propagation, giving rise to plausible alignment performance even in unconstrained environments.

Problem Formulation: Suppose we have N training face samples denoted by $\boldsymbol{I}$, we let $\boldsymbol{p} = [p_1, p_2, \cdots, p_L] \in \mathcal{P} \in \mathbb{R}^{2\times L}$ denote L heat maps with L points, where $p_i = p_i(j,k)$ represents the horizontal j and vertical k coordinates for the i0-th landmark, the vector $\boldsymbol{p}^* = [p_1^*, \cdots, p_L^*]$ denote the groundtruth maps, respectively. The main goal of our model is to minimize the discrepancy between the landmark maps and groundtruth as follows:

$$J = \frac{1}{N}\sum_{i=1}^{N}\sum_{j,k} \| f_{\text{DHGN}}(\boldsymbol{I}_i) - \boldsymbol{p}_i(j,k)\|_2^2 \tag{1}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm to measure the heat map residuals and f_{DHGN} provides the network parameters of the proposed DHGN network architecture including deformable transformer and hourglass modules, respectively.

Deformable Transformer Module: The objective of our model is to learn the network parameters in f_{DHGN} via (1). Motivated by the intuitions from [17, 18], we develop a deformable transformer module, which requires to learn a set of transformation offsets and then refine the positions of sampling mask without any supervision signals. With the deformable masking, the attentional facial-part features are exploited for robust face alignment. Specially, our module typically consists of a deformable regression subnet and a deformable masking operator. The regression subnet is designed to produce the offset values, while the masking operator aims to filter the attentional regions directly from the input maps and then feed them to the backbone alignment networks.

As demonstrated in Fig. 2, the input of the proposed deformable transformation is $\boldsymbol{U}$ which contains the same size with the face input. Our regression subnet is fed with feature map $\boldsymbol{U}$ and then directly estimate a set of offsets denoted by $\Delta\boldsymbol{p}$ with the dimension of L landmarks for deformable masking. Obviously, we noted that the regression subnet is differential and the offsets is leveraged to refine the positions of the initialized mask $\boldsymbol{p}^0$, *i.e.* statistical mean shape, by adding $\Delta\boldsymbol{p}$. Having obtain the refined mask located at $\boldsymbol{p}^0 + \Delta\boldsymbol{p}$, our designed masking operator crops the local patches based on the new mask and then feed them to the deformable CNN. As a result, the output feature maps V are explored with the deformable receptive fields by following [18], rather than using the conventional CNN with fix geometric structures.

Suppose we have these cropped patches denoted by $\mathbf{x}^t(\mathbf{p}+d)$, where d is the patch size that was specified to 26 in our experiments. To perform an end-to-end optimization procedure, we also provide the derivatives of the shape with respect to the loss, which is computed for each landmark p (certain point from the shape $\mathbf{p}$) as follows:

$$\frac{\partial J}{\partial p} = \frac{\partial \mathbf{x}}{\partial p}\frac{\partial J}{\partial \mathbf{x}}, \tag{2}$$

$$\frac{\partial \mathbf{x}}{\partial p} = \nabla(\mathbf{x}(p+d)), \tag{3}$$

where d is the size of sampled shape-index patches, $\mathbf{x}(p)$ denotes the pixel value located at the landmark p and ∇ denotes the gradient-image w.r.t the cropped image patch, respectively. Since the derivatives of the shape-image are not strictly differentiable for 2D images, the value is approximated by the gradient of the image. Specifically, $\nabla(\mathbf{x}(p+d))$ is calculated by the Sobel operator in size of $d\times d$ which is convolved on the image patches. The final result is summed up by performing gradients of total landmarks.

During training procedure, the network parameters of the deformable transformer are learned by using standard back-propagation and alignment loss function (1). In another sense, both the regression subnet and masking operator is approximately differential and can be readily integrates in existing alignment networks (hourglass networks [16] module is employed in our work) for end-to-end feature learning.

Backbone Hourglass Networks: Generally, there exist a number of possible solutions which can be used to exploit the regression functions for face alignment. In our work, we exploit the widely-used hourglass networks as the backbone alignment network. Specifically, the hourglass networks are passed across different scales and consolidated to best capture various spatial relationships associated with facial landmarks, where the multi-scale and spatial attentional information are simultaneously exploited for robust feature extraction. Compared with traditional coordinates-based regression, the hourglass modules leverage a final set of predictions of heat maps to reduce the variations of the regression target. To further improve the capacity of the deep regression model, we expand on a single hourglass by consecutively placing multiple hourglass modules together in an end-to-end manner [16].

Inference of face alignment: Having obtained the output resolution of our DHGN network, two consecutive rounds of 1×1 convolutions are applied to produce the final network predictions. The output of the network is a set of heat maps, where the map typically predicts the probability of each facial landmark's presence at every pixel. Hence, we regard the positions by maximizing the responses as the resulting coordinates for fine-grained facial landmark localization.

3. EXPERIMENTS

To evaluate the effectiveness of the proposed method, we conducted folds of experiments on the standard benchmarks.

Evaluation Datasets and Protocol. *300-W [19]*: This dataset is composed by several sub-datasets for training and testing, *i.e.*, LFPW [20], HELEN [21], AFW [1], XM2VTS [22] and IBUG [23], where the 300-W organizer provides 68-pts annotations. By strictly following the 300-W protocol, we trained our model with the LFPW training set, the HELEN training set, the AFW dataset and tested it on the LFPW testing set, the HELEN testing set and the IBUG dataset, respectively. Moreover, we investigated our approaches on another dataset setting, where we consider the testing sam-

Table 1. Comparisons of averaged errors (100%) normalized by inter-pupil-distance of our DHGN with existing methods *in chronological order* on the 300-W (68-lm) Commonset, Challengingset and Fullset. Compared with these methods, our model achieves very competitive performance.

Method	Commonset	Challengingset	Fullset
FPLL [1]	8.22	18.33	10.20
DRMF [25]	6.65	19.79	9.22
RCPR [24]	6.18	17.26	8.35
SDM [7]	5.57	15.40	7.50
LBF [14]	4.95	11.98	6.32
CFAN [9]	5.50	16.78	7.69
CFSS [8]	4.73	9.98	5.76
TCDCN [26]	4.80	8.60	5.54
RAR [27]	4.12	8.35	4.94
MDM [13]	4.83	10.14	5.88
DSSD [28]	4.16	9.20	5.59
DeepReg [29]	4.36	7.42	4.96
TCD [10]	3.67	7.62	4.44
CPR [30]	3.39	8.14	4.36
DHGN	**3.38**	**6.23**	**3.95**

ples from the LFPW and HELEN datasets as the Commonset and the 135-image IBUG dataset as the Challengingset, and the union of them as the 689-image Fullset. *COFW* [24]: The Caltech Occluded Face in the Wild (COFW) dataset consists of 1345 training face images and 507 testing face images, which were collected from the Internet. All face images are annotated with 29 landmarks together with the visibility/invisibility information. We conducted experiments and evaluated our methods only on its testing set. Note that our model was trained on the training samples from the 300-W dataset, without using any training images from the COFW dataset. For the evaluation protocol, we employed the root mean squared error (RMSE), where the point-to-point discrepancy is normalized by the inter-ocular distance. Besides, we averaged the RMSEs of testing frames within each division and then average them as the final performance.

Implementation Details. For the input data preparation, we detected faces on the whole dataset by enlarging the groundtruth annotations. Then we rescaled both the detected facial images with padding zeros and the corresponding annotations with the restricted output scales 64×64. The input image first passes through max pooling in the size of 2×2. Then the downscaled features are processed by a bottom-up layers including three convolutional layers with kernels [128, 128, 256]. The following layers is equipped with a symmetric top-down structure. In addition, a residual module is deployed for multi-scale message passing across layers. In the term of the stacking hyper-parameter of our model, we found using four hourglass modules are sufficient for high-performance alignment performance.

Results and Analysis. We compared our proposed D-HGN with the state-of-the-art face alignment methods and Table 1 tabulates the comparisons of averaged errors of our method compared with the state-of-the-arts on 300-W dataset. From these results, we see that our proposed DHGN significantly outperforms other face alignment methods by a large margin, which is because our designed deformable modules exploit more cues to learning discriminative features for robust face alignment. We also evaluated our approach on the COFW dataset regarding of occlusions and the results are tabulated in Table 2. From the results, we achieve very compelling performance on the challenging cases due to large poses, diverse expressions and severe occlusions. This also proves the effectiveness of the proposed deformable masking method, where these learned shape-sensitive cues are helpful to promote the alignment performance.

Table 2. Comparison of the averaged errors and the failure rates (threshold at 0.08%) on the COFW dataset. Our DHGN significantly achieves robustness to severe partial occlusions.

Method	Averaged Error	Failure Rate (%)
FPLL [1]	8.79	38.46
ESR [6]	11.20	36.00
RCPR [24]	8.50	20.00
HPM [31]	7.46	13.24
RPP [32]	7.52	16.20
SDM [7]	8.77	24.32
CFAN [9]	8.38	19.14
TCDCN [26]	8.05	15.31
DSSD [28]	6.17	8.23
DHGN	**5.29**	**6.94**

4. CONCLUSION

In this paper, we have proposed a deformable hourglass networks (DHGN) approach for robust face alignment. The experimental results have demonstrated the effectiveness of the proposed approach on several widely-evaluated face alignment datasets. One desirable direction of this work is to restrict the testing time tolerance and to plug some runtime deep compression strategies in the meantime of the shape updates, *e.g.*, by following the runtime network pruning in the future works.

5. REFERENCES

[1] Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR. (2012) 2879–2886

[2] Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: ICCV. (2015) 2983–2991

[3] Liu, F., Zeng, D., Zhao, Q., Liu, X.: Joint face alignment and 3d face reconstruction. In: ECCV. (2016) 545–560

[4] Liu, H., Lu, J., Feng, J., Zhou, J.: Two-stream transformer networks for video-based face alignment. TPAMI **40**(11) (2018) 2546–2554

[5] Bouwmans, T., Javed, S., Zhang, H., Lin, Z., Otazo, R.: On the applications of robust PCA in image and video processing. Proceedings of the IEEE **106**(8) (2018) 1427–1457

[6] Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR. (2012) 2887–2894

[7] Xiong, X., la Torre, F.D.: Supervised descent method and its applications to face alignment. In: CVPR. (2013) 532–539

[8] Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: CVPR. (2015) 4998–5006

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

[10] Kumar, A., Chellappa, R.: Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. CVPR (2018)

[11] Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR. (2018)

[12] Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: A convolutional neural network for robust face alignment. In: CVPR. (2017) 2034–2043

[13] Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In: CVPR. (2016) 4177–4187

[14] Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 FPS via regressing local binary features. In: CVPR. (2014) 1685–1692

[15] Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: CVPR Workshop. (2017) 2025–2033

[16] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. (2016) 483–499

[17] Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NIPS. (2015) 2017–2025

[18] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV. (2017) 764–773

[19] : 300 faces in-the-wild challenge. `http://ibug.doc.ic.ac.uk/resources/300-W/`

[20] Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR. (2011) 545–552

[21] Le, V., Brandt, J., Lin, Z., Bourdev, L.D., Huang, T.S.: Interactive facial feature localization. In: ECCV. (2012) 679–692

[22] Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: The extended m2vts database. AVBPA **964** (1999) 965–966

[23] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCVW. (2013)

[24] Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: ICCV. (2013) 1513–1520

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

[26] Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. TPAMI **38**(5) (2016) 918–930

[27] Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.A.: Robust facial landmark detection via recurrent attentive-refinement networks. In: ECCV. (2016) 57–72

[28] Liu, H., Lu, J., Feng, J., Zhou, J.: Learning deep sharable and structural detectors for face alignment. TIP **26**(4) (2017) 1666–1678

[29] Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: CVPR. (July 2017)

[30] Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In: CVPR. (2018)

[31] Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In: CVPR. (2014) 1899–1906

[32] Yang, H., He, X., Jia, X., Patras, I.: Robust face alignment under occlusion via regional predictive power estimation. TIP **24**(8) (2015) 2393–2403