Unconstrained Face Alignment without Face Detection

Xiaohu Shao[1,2], Junliang Xing[3], Jiangjing Lv[1,2], Chunlin Xiao[4],
Pengcheng Liu[1], Youji Feng[1], Cheng Cheng[1]
[1] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Institute of Automation, Chinese Academy of Sciences
[4] CloudWalk Technology
{shaoxiaohu,lvjiangjing,liupengcheng,fengyouji,chengcheng}@cigit.ac.cn
jlxing@nlpr.ia.ac.cn   xiaochunlin@cloudwalk.cn

Abstract

This paper introduces our submission to the 2nd Facial Landmark Localisation Competition. We present a deep architecture to directly detect facial landmarks without using face detection as an initialization. The architecture consists of two stages, a Basic Landmark Prediction Stage and a Whole Landmark Regression Stage. At the former stage, given an input image, the basic landmarks of all faces are detected by a sub-network of landmark heatmap and affinity field prediction. At the latter stage, the coarse canonical face and the pose can be generated by a Pose Splitting Layer based on the visible basic landmarks. According to its pose, each canonical state is distributed to the corresponding branch of the shape regression sub-networks for the whole landmark detection. Experimental results show that our method obtains promising results on the 300-W dataset, and achieves superior performances over the baselines of the semi-frontal and the profile categories in this competition.

1. Introduction

Face alignment, which is to locate predefined facial landmarks on images given face detection results, is one of the most important tasks in the field of computer vision. Many researchers have devoted great efforts to solving this task, and recently regression based algorithms have become the dominant solution for the face alignment task [6, 7, 8, 32, 23, 24, 2, 14, 10, 22, 27, 30, 31] because of their high precision and efficiency. Currently, as deep learning introduced to face alignment, many promising deep learning based methods [27, 19, 17, 29, 21, 13] have also been developed to further improve performances.

Although current alignment methods have achieved nearly perfect results on (near) frontal images based on proper face detectors, they are still facing two main challenges: (1) Heavily dependent on initialization of face detectors, if the face detector in test phase provides an improper face rectangle, or fails in detection on a face image, the performance of subsequent face alignment would degrade a lot. (2) Alignment for faces with arbitrary poses, *e.g.*, faces with yaw angle larger than 45° is not satisfied because of unsatisfied face detection, insufficient training samples, and the lack of research attentions.

In order to boost research in face alignment addressing the above challenges, the 2nd Facial Landmark Localisation Competition [26] - the Menpo Benchmark is held in conjunction with CVPR 2017 . We take part in both the frontal and profile categories and achieve better performances over the baselines. We present a deep architecture to directly detect facial landmarks on faces with arbitrary poses. Specifically, different from traditional work which only focus on detecting landmarks based on face detection results [8, 23, 22, 24, 2, 14, 10, 27, 30, 31, 27, 19, 17, 20], or work which joint face detection and face alignment [5, 29, 13, 28], our method does not adopt any face detector and directly detect facial landmarks on an image in a bottom-up manner. This advantage is very suitable for this competition, as the training and testing data sets of all categories do not provide any face boxes and there are many profile images are difficult to be detected by traditional face detectors.

The proposed architecture without face detection, whose framework is shown in Figure. 1, mainly consists of two stages, a Basic Landmark Prediction Stage and a Whole Landmark Regression Stage. The basic landmarks of all faces, *e.g.*, landmarks of the centers of two pupils, nose tip and mouth corners, are detected by a sub-network of landmark heatmap and affinity field prediction. This step, in-

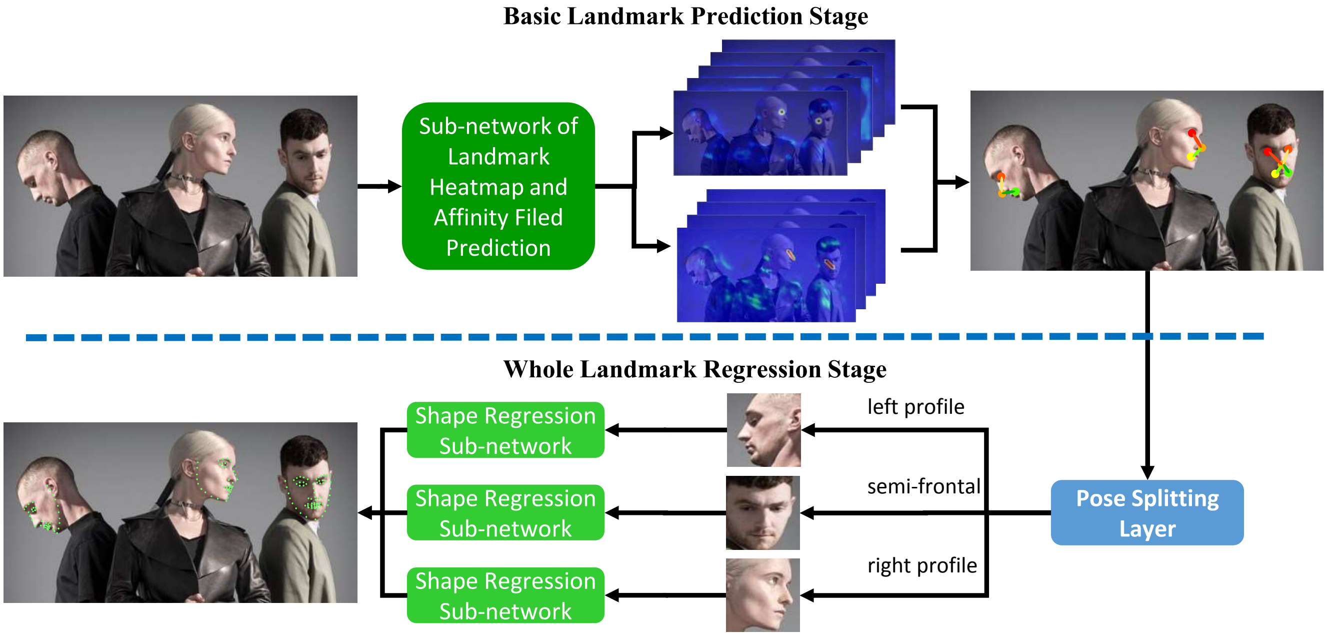

Basic Landmark Prediction Stage
Sub-network of Landmark Heatmap and Affinity Filed Prediction
Whole Landmark Regression Stage
Shape Regression Sub-network
Shape Regression Sub-network
Shape Regression Sub-network
left profile
semi-frontal
right profile
Pose Splitting Layer

Figure 1. The pipeline of the proposed deep regression architecture.

spired by work [3], allows the proposed network directly detect landmarks on arbitrary images from end to end, which is different from the traditional face alignment methods using face detection as an initialization. At the latter stage, the coarse canonical state and pose of each face can be generated by a Pose Splitting Layer based on the positions of basic landmarks. The pose of each face, *e.g.*, semi-frontal, left profile, right profile, can be calculated by a simple geometry principle of the basic landmarks. According to its pose, each canonical state is distributed to the corresponding branch of the shape regression sub-network for the whole landmark detection. Our unconstrained face alignment without using face detection as initialization not only obtains promising results on the 300-W dataset [16] compared with many state-of-the-art methods, but also achieves superior performance over the baseline in both the semi-frontal and profile categories of the Menpo Benchmark.

The end-to-end face alignment reduces adverse effects from unsatisfied face detection results on faces of arbitrary poses, its time complexity of basic landmark prediction is constant to number of faces in an image. In the future, by employing optimization of implementation, it will improve the efficiency of face alignment for multiple faces in an image because its bottom-up detection way.

2. Our Method

This section describes the framework (shown in Figure 1) and details of the two stages in our proposed deep architecture. In the following, we elaborate on the designations of the two stages, the Basic Landmark Prediction Stage (BLPS) and the Whole Landmark Regression Stage (WLRS), and then introduce implementation details of the whole model.

Given an input image $I \in \Re^{w \times h}$ (w and h are the width and height of the image, respectively), the objective of face alignment is to locate the predefined shape $\mathbf{S} = [\mathbf{x}_1, ..., \mathbf{x}_n]^T \in \Re^{2 \times n}$ with n positions of landmarks $\mathbf{x} = (x, y)^T$ on the faces in the image.

2.1. Basic Landmark Prediction Stage

In this stage, the basic landmarks $\mathbf{S}_{basic}$ instead of face rectangles of all faces on I are predicted firstly. We choose five landmarks, the landmarks of the two centers of pupils, nose tip, two mouth corners as the basic landmarks, for they are more saliency and easier to be detected than other landmarks. In order to detect $\mathbf{S}_{basic}$ on each face, we explore a module in which heatmaps $\mathbf{H}$ responsed by all the landmarks and association fields $\mathbf{L}$ between two associated landmarks are detected via a sub-network of landmark heatmap and affinity field prediction (see in Figure 2). This module is motivated by the work of Part Affinity Fields for Part Association (PAF) in a bottom-up way [3]. Different from PAF for pose estimation with multiple stages, the sub-network used in our architecture is designed for responses of positions and associations of facial landmarks with only one stage.

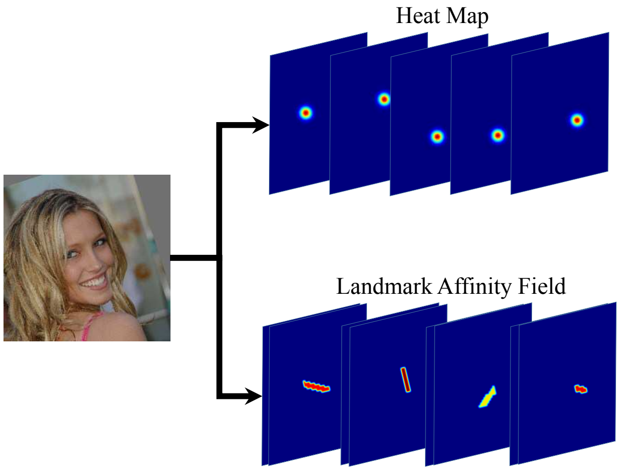


Figure 2. The heatmaps for the basic landmarks.

2.1.1 Landmark Heatmap

Heatmaps responsed by landmarks of $\mathbf{S}_{basic}$ are predicted in this stage. The ground truth value of point $\mathbf{p}$ on heatmap $\mathbf{H}^*_{i,k} \in \Re^{w\times h}$ for the i^{th} ($i \in \{1, 2, ..., n\}$) landmark $\mathbf{x}_i$ and the k^{th} face can be formulated by Guassian peaks as:

$$\mathbf{H}^*_{i,k}(\mathbf{p}) = \exp(-\frac{||\mathbf{p} - \mathbf{x}_{i,k}||_2^2}{\lambda}), \tag{1}$$

where λ is a positive value proportional to the face size. Each channel of the heatmap for each landmark of all faces on I can be calculated by the following equation:

$$\mathbf{H}^*_i(\mathbf{p}) = \max_k \mathbf{H_{i,k}}(\mathbf{p}). \tag{2}$$

After these heatmaps generated, all the positions of landmarks for all faces is located firstly, however, because of the lack of constraint of the whole shape, there are many false alarms in these heatmaps and how to assemble them to form whole faces is still a problem.

2.1.2 Landmark Affinity Field

Landmarks affinity field prediction is introduced to reduce false alarms of heatmaps and deploy a solution to assemble candidate landmarks. Similar to PAF, we define the landmark affinity vector field for describing the association between any two landmarks. The part connected by arbitrary landmarks $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ is used to represent association of the two landmarks. Suppose there are C types of parts for each face, the ground truth value at point $\mathbf{p}$ of the c^{th} landmark affinity vector field $\mathbf{L}$ for the k^{th} face is denoted as:

$$\mathbf{L}^*_{c,k}(\mathbf{p}) = \begin{cases} \mathbf{v} & \text{if } \mathbf{p} \text{ on part } c,\, k \\ \mathbf{0} & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{v}$ is the unit vector in the direction of the part $\mathbf{L}^*_{c,k}$. Whether $\mathbf{p}$ is on the part or not is determined by a distance threshold of the line connected by $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$. The value of landmark affinity field $\mathbf{L}^*_c(\mathbf{p}) \in \Re^{w\times h\times 2}$ of type c for all faces is the average of all $\mathbf{L}^*_{c,k}(\mathbf{p})$ where different faces' parts overlap.

The association score E_c of each part can be measured by integrating values $\mathbf{L}^*_c(\mathbf{p})$ along the part. As the maximum of types $C = \mathrm{C}_5^2$ brings a big inference burden to find all the candidate landmarks that can be connected belong the same face, we only consider the parts which are assembled by the landmark of nose tip connected to other four type landmarks, so that $c = 4$ in our framework. For the c^{th} part, the optimization goal to find a matching with maximum scores for all possible connections Z_c among all corresponding landmarks:

$$\max_{Z_c} E_c = \max_{Z_c} \sum_{m\in\Delta_{i_1}} \sum_{n\in\Delta_{i_2}} E_{mn}, \tag{4}$$

where Δ_{i_1} is the set of all candidate nose landmarks, Δ_{i_2} belongs to the set of all candidate of other type landmarks.

2.1.3 Sub-network Learning

We design a CNN network derived from the VGG-19 [18] model, which is shown in Figure 3, for jointly learning heatmaps $\mathbf{H}$ and affinity fields $\mathbf{L}$ of landmarks on an arbitrary image I. The whole sub-network has 14 convolution layers, with the first 8 convolution layers initialized by VGG 19 model trained on ImageNet classification dataset. Two sibling branches are followed from the output of conv3_4, where the first branch is used for the task of predicting the heatmap $\hat{\mathbf{H}}$ of all basic landmarks and the other one is used for another task of predicting associated facial fields $\hat{\mathbf{L}}$ of all parts. The loss functions for the above two tasks are:

$$\mathcal{L}_\mathbf{H} = \sum_{i=1}^{n} ||\hat{\mathbf{H}} - \mathbf{H}^*||_2^2, \tag{5}$$

$$\mathcal{L}_\mathbf{L} = \sum_{c=1}^{C} ||\hat{\mathbf{L}} - \mathbf{L}^*||_2^2. \tag{6}$$

Instead of using two separated branches for the two tasks learning and multiple stages for repeatedly prediction in work [3], we find that it already gets a satisfied performance for basic landmarks detection by using only one stage. Landmarks belong to a whole face can be drawn on an image after all the basic landmarks and affinity parts are predicted. Example results of the first stage are shown in Figure 4. Usually, full basic landmarks are detected for the semi-frontal faces, there are one or two landmarks are missing predicted because of self-occlusion for the profile faces.

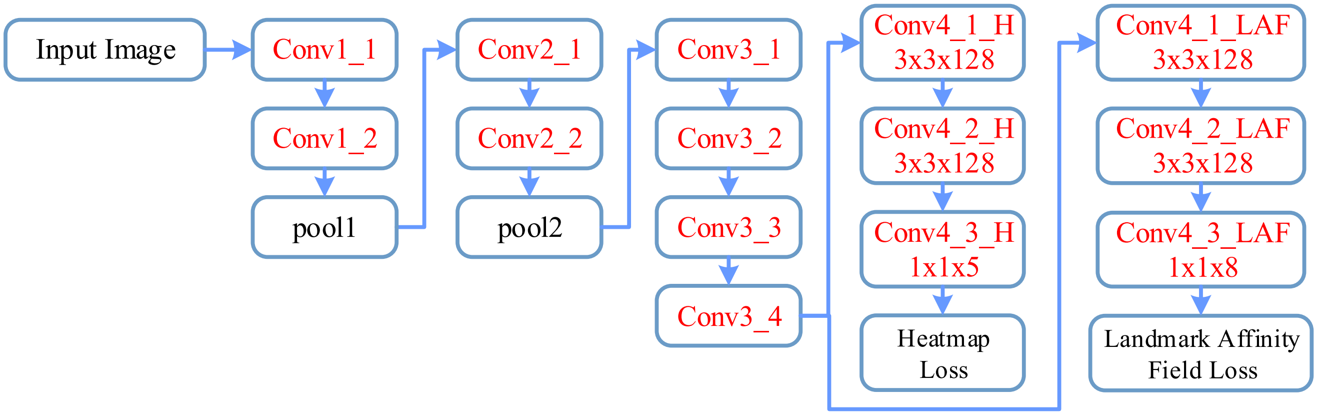


Figure 3. The architecture of the sub-network of lanmdark heatmap and affinity field prediction.

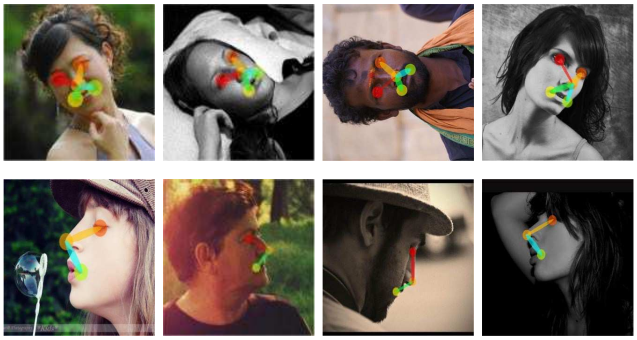


Figure 4. Results of the landmark heatmap and affinity field sub-network, the top and the bottom line show detection results on semi-frontal and profile faces, respectively. Landmarks occluded because of large pose on profile faces would be labeled invisible by the sub-network.

2.2. Whole Landmark Regression Stage

2.2.1 Pose Splitting Layer

According to the visibility of landmarks in each face, faces are automatically divided into three different pose types (left profile, right profile and semi-frontal). In general, the overall procedure of pose splitting is summarized in Algorithm 1, where p_{LE}, p_{RE}, p_N, p_{LM}, and p_{RM} represent locations of left eye center, right eye center, nose tip, left mouth corner, and right mouth corner, respectively. Especially for the profile category in the Menpo Benchmark, Algorithm 2 is introduced where there are only two poses to be classified, left profile and right profile.

After each face is classified into specific pose type, three different predefined canonical templates are used to normalize faces with three pose, respectively, as shown in Figure 5. Specifically, we use the following formula to define the similarity transformation.

$$\begin{bmatrix} \bar{x}_i \\ \bar{y}_i \end{bmatrix} = \begin{bmatrix} a & b & t_x \\ -b & a & t_y \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \qquad (7)$$

where x_i, y_i are the detected landmarks, $\bar{x}_i$, $\bar{x}_i$ are the pre-defined canonical positions of templates, a, b, t_x and t_y are the parameters of similarity transformation which can be calculated by the canonical templates and predicted basic landmarks. For each pose type, each face has at least two visible landmarks, they are enough for pose calculation.

Algorithm 1 Pose Splitting.

```
Input: $landmarks = \{p_{LE}, p_{RE}, p_N, p_{LM}, p_{RM}\}$
Output: $pose\_type$ $(0 \to semi - frontal, 1 \to$
    $left\_profile, 2 \to right\_profile)$
function PoseSplitting($landmarks$)
    $result \leftarrow 0$
    if $p_{LE} == invisble\ and\ p_{LM} == invisble$ then
        $result \leftarrow 1$
    end if
    if $p_{RE} == invisble\ and\ p_{RM} == invisble$ then
        $result \leftarrow 2$
    end if
    return $result$
end function
```

2.2.2 Shape Regression Sub-network

As the landmarks of faces with different poses in the Menpo Benchmark are differently defined, $n = 68$ for semi-frontal face labeling, $n = 39$ for left and right profile faces labeling in the Menpo Benchmark, we explore three branches of CNNs for the three poses respectively. With the canonical face image F of the corresponding pose input, a shape regression sub-network is explored to learn positions of shape

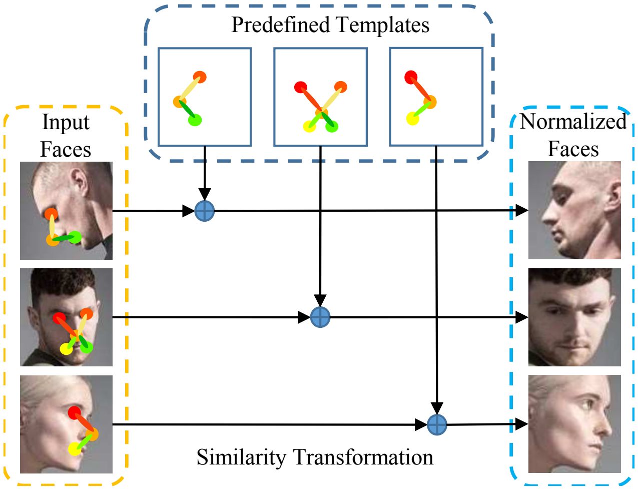


Figure 5. The designation of the Pose Splitting Layer in our deep architecture, which automatically normalize the face images from different poses, and enables the whole model to be trained from end to end.

S^* with whole landmarks labeled in the training set. This sub-network can be built based on a traditional deep regression network for face alignment. Standard Euclidean distance between the ground truth $\mathbf{S}^*$ and predicted landmarks $\hat{\mathbf{S}}$ is used as the loss of the final landmark detection:

$$\mathcal{L}_{\mathbf{S}} = ||\hat{\mathbf{S}} - \mathbf{S}^*||_2^2. \quad (8)$$

After $\hat{S}$ is predicted, it would be projected back into the coordinate space of the initial image I by using the inverse transformation matrix calculated by the Pose Splitting Layer.

2.3. Implementation Details

There are mainly two steps of our whole architecture training, the training for the sub-network of landmark heatmap and affinity field prediction and the shape regression sub-networks, and there is no parameters to be learned in the Pose Splitting Layer. During the former training, the patches are cropped from images with arbitrary sizes and resized to 80 × 80 resolution with a roughly 40 × 40 face. The loss weights of $\mathcal{L}_{\mathbf{H}}$ and $\mathcal{L}_{\mathbf{L}}$ are set to the same value of 1. In order to improve the robustness of each shape regression sub-network, a few of extension samples are generated for each input face image with 448 × 448 resolution by disturbing face regions by proper translation, scaling, rotation and flipping. The branch of sub-network for left/right direction profile faces is built based on the VGG-S network [4], which comprises eight learnable layers, five among them are convolutional and the last three are fully-connected. We modify the output of last layer from 1000 to 2 × 39 for predicting the 39 landmark positions. The branch of sub-network for semi-frontal faces is built on the network of [12] for its invariant to various initialization brought by input face regions.

Algorithm 2 Pose Splitting for Profile Category.

```
Input: $landmarks = \{p_{LE}, p_{RE}, p_N, p_{LM}, p_{RM}\}$
Output: $pose\_type$ ($0 \rightarrow semi\_frontal, 1 \rightarrow$
    $left\_profile, 2 \rightarrow right\_profile$)
function PoseSplittingProfile($landmarks$)
    $result \leftarrow 0$
    if $p_{LE} == invisible$ or $p_{LM} == invisible$ then
        $result \leftarrow 1$
    else if $p_{LE} == invisible$ or $p_{LM} == invisible\}$ then
        $result \leftarrow 2$
    else
        if $abs(norm(p_{LE} - p_N))+$
        $abs(norm(p_{LM} - p_N)) < abs(norm(p_{RE} - p_N))+$
        $abs(norm(p_{RM} - p_N))$ then
            $result \leftarrow 1$
        else
            $result \leftarrow 2$
        end if
    end if
    return $result$
end function
```

3. Experiments

We first introduce our experimental settings during our training and test steps. Then compare our method with other state-of-the-art methods on the 300-W dataset. Finally, we show the comparison between our method and the baseline of the Menpo Benchmark. The whole architecture is implemented using the Caffe software package [9].

3.1. Training Datasets and Experimental Settings

In order to prove the effectiveness or our approach, we use the three following datasets for training, validation and test:

CelebA [11]: CelebFaces dataset contains 202,599 images and each image contains only one face. It covers large pose variations and background clutter. This dataset contains faces with similarly pose distribution of Menpo dataset, and it provides 5 basic landmarks for each face, so it is very suitable for the training of our sub-network of landmark heatmap and affinity field prediction to predict the basic landmarks.

300-W [16]: The dataset consists re-annotated five existing datasets with 68 landmarks: iBug, LFPW, AFW, HELEN, and XM2VTS. We follow the work [30] to use 3, 148 images for training and 689 images for testing. The testing set is spitted into three parts: common subset (554 images),

Table 1. The partition of the whole Menpo Data.

	semi-frontal	profile
Training	6010	2069
Validation	669	231
Test	12006	4253

challenging subset (135 images) and the full set (689 images). The model trained on this training subset is used for comparison with other state-of-the-art methods. By fine-tuning on the Menpo Benchmark, we get all the models of shape regression networks.

Menpo Data: This dataset are separated into two categories, the semi-frontal subset and the profile subset. The former is annotated with the standard 68 point landmarks following the principle of [16], and the latter has symmetric 39 landmarks for the left and right profile faces. The partition of the whole dataset in our experiments is described in the Table 1. Among them, the training data and validation data are separated randomly with the ration of 10:1 on the released training data with ground truth landmarks. The released test data has no public ground truth. In addition, there is no face rectangle labeled for any face in this dataset. Each branch of our shape regression sub-networks is trained on the subset of training dataset, respectively.

We use the normalized mean error (NME) to evaluate performance of different methods. For semi-frontal face with 68 landmarks annotated, inter-ocular distance is employed to normalize mean error on 300-W following work [15]. Because there is no released ground truth for test subset of Menpo Data, it is unclear of the normalization way for profile category in this competition.

3.2. Comparison with the State-of-the-art Face Alignment Methods

It is not easy to compare our method with other work of face alignment for profile faces currently for the variety of annotations. In this section, we only compare our approach with recently proposed methods [23, 24, 2, 10, 15, 1, 30, 25] on the 300-W dataset, see in Table 2. Because all the data of 300-W are all 68-landmark labeled, faces output by the Basic Landmark Prediction Stage are only distributed to the branch of shape regression sub-network for semi-frontal face. For each image, landmarks of multiple faces may be detected by our method, their overlapping areas with the face box provided by the 300-W dataset is used as a guide for the final landmark evaluation. There 23 faces failed to be detected by BLPS, and the corresponding official boxes of 300-W are used as supplement for further face alignment. The results show that the Basic Landmark Prediction Stage of the proposed architecture is able to provide a good initialization for further landmark detection even without face detection. By connecting with a strong shape regression sub-network, our method gets better performance than other methods.

Table 2. The performance of our proposed method compared with other methods on the 300-W dataset.

Method	Common Subset	Challenging Subset	Full Set
RCPR [1]	6.18	17.26	8.35
SDM [23]	5.57	15.40	7.52
ESR [2]	5.28	17.00	7.58
CFAN [27]	5.50	16.78	7.69
DeepReg [17]	4.51	13.80	6.31
LBF [14]	4.95	11.98	6.32
TCDCN [29]	4.80	8.60	5.54
CFSS [30]	4.73	9.98	5.76
DDN [25]	-	-	5.59
Proposed	4.45	8.03	5.15

3.3. Comparison with the Baseline Method of the Menpo Benchmark

As the categories of semi-frontal and profile faces are separated, we do a necessary modify on our Pose Splitting Layer. For the evaluation on semi-frontal faces, all the faces are only input to the branch of regression sub-network for frontal faces, which is mentioned in 3.2. For the evaluation on profile faces, the output of Pose Splitting Layer for frontal faces is disabled. In order to evaluate the effectiveness of our BLPS method, we firstly evaluate it on the validation subsets, there are 6 of 699 faces in the semi-frontal category and 8 of 231 faces in the profile category not detected. When the basic landmarks are detected, the Pose Splitting Layer achieves an average accuracy of 99% to classify the left and right profile faces in the validation subset of the profile category. In the testing set, there are 10 and 134 miss detections in the two categories, respectively. For these samples, the whole images are supposed as input for further face alignment. Because there are a few of cases that multiple faces are detected in an image, the alignment result of the most central face is selected for the evaluation. The Cumulative Error Distribution (CED) curves of our method and Menpo baselines, which are evaluated by the organizers, are illustrated in Figure 7.

4. Conclusion and Future Work

In this paper, we present an architecture for face alignment without using face detection as an initialization, and describe our submission to the 2nd Facial Landmark Localisation Competition. For the initialization task of face alignment, the prediction of heat maps of basic landmarks and the association fields of landmark-pair takes place of face regions detection. By cooperating with the Pose Split-

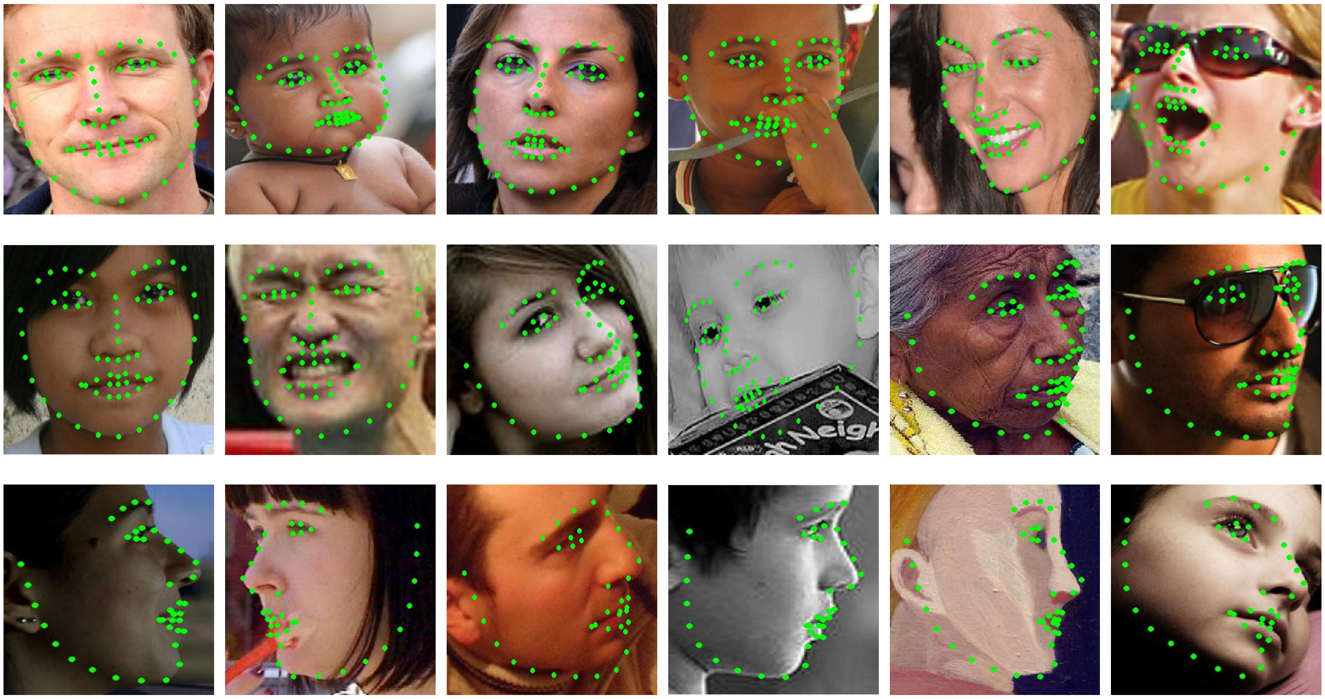
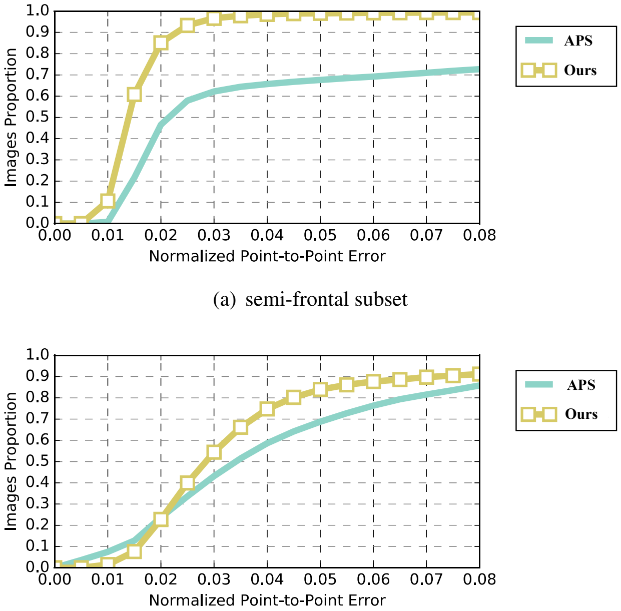


Figure 6. Examples of the face alignment results produced using the proposed method. The images show the results on the 300-W dataset, semi-frontal subset and profile subset of the Menpo benchmark from the top row to the bottom row.



(a) semi-frontal subset

(b) profile subset

Figure 7. The CEDs using the 68 landmarks for (a) semi-frontal category and 39 landmarks for (b) profile category. APS is the baseline of Menpo Benchmark.

ting Layer and branches of shape regression sub-network for each pose, the whole landmarks on faces with arbitrary poses are detected accurately. Our proposed method without depending any face boxes not only obtains promising results on the 300-W dataset, but also outperforms the baselines of the Menpo Benchmark.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, abs/1405.3531, 2014.

[5] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014.

[6] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[8] P. Dollar, P. Welinder, and P. Perona. Cascaded pose regression. *Proceedings of IEEE International Conference on Computer Vision*, 238(6):1078–1085, 2010.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[10] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[12] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage reinitialization for high performance facial landmark detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, accepted, 2017.

[13] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.

[14] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[15] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016.

[16] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2013.

[17] B. Shi, X. Bai, W. Liu, and J. Wang. Deep regression for face alignment. *arXiv preprint arXiv:1409.5230*, 2014.

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

[19] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[20] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[21] Y. Wu and T. Hassner. Facial landmark detection with tweaked convolutional neural networks. *arXiv preprint arXiv:1511.04031*, 2015.

[22] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan. Towards multi-view and partially-occluded face alignment. In *Proceedings of IEEE International Conference on Computer Vision*, 2014.

[23] X. Xiong and F. Torre. Supervised descent method and its applications to face alignment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[24] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[25] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. *arXiv preprint arXiv:1605.01014*, 2016.

[26] S. Zafeiriou, G. Trigeorgis, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step closer to the solution. *CVPRW*, 2017.

[27] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Proceedings of European Conference on Computer Vision*, 2014.

[28] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proceedings of European Conference on Computer Vision and Pattern Recognition*. 2014.

[30] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[31] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[32] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.