Stacked Hourglass Network for Robust Facial Landmark Localisation

Jing Yang
y.jing2016@gmail.com

Qingshan Liu
qsliu@nuist.edu.cn

Kaihua Zhang
zhkhua@gmail.com

Nanjing University of Information Science and Technology
Nanjing, China

Abstract

With the increasing number of public available training data for face alignment, the regression-based methods attracted much attention and have become the dominant methods to solve this problem. There are two main factors, the variance of the regression target and the capacity of regression model, affecting the performance of the regression task. In this paper, we present a Stacked Hourglass Network for robust facial landmark localisation. We first adopt a supervised face transformation to remove the translation, scale and rotation variation of each face, in order to reduce the variance of the regression target. Then we employ a deep convolutional neural network named Stacked Hourglass Network to increase the capacity of the regression model. To better evaluate the proposed method, we reimplement two popular cascade shape regression models, SDM and LBF, for comparison. Extensive experiments on four challenging datasets, COFW, IBUG, 300W and the Menpo Benchmark, prove the effectiveness of the proposed method.

1. Introduction

A key step towards understanding people in images and videos is accurate facial landmark localisation [20, 7, 44, 37, 39, 29, 1], also known as face alignment. Given a single RGB face image, face alignment algorithms try to determine the precise pixel location of crucial points on the face. Achieving an efficient and effective system to locate facial landmarks is beneficial to higher level tasks, *e.g.* facial attribute analysis [25], expression analysis [28], and face recognition [37, 46, 38, 21, 22, 34]. Face alignment also serves as a fundamental tool in field such as human-computer interaction and animation [6]. Although the past decades have witnessed intensive research and great strides in designing and developing algorithms for face alignment [15, 11, 26, 14, 7, 12, 8, 29, 13, 24], it remains a challenging task, especially when the face images suffer from large pose variations and partial occlusions.

Recently, the cascade shape regression method has become the mainstream approach for robust face alignment, and it is primarily motivated by the cascade pose regression presented in [14], which attempts to directly infer a face pose through a discriminative non-linear mapping between the textual features and the object's pose parameters. It has inspired a volume of research works in the field of face alignment [7, 12, 39, 29], in which a mapping function is directly learned from the image appearance to the face shape. The discriminations among these approaches mainly lie in the regression methods and the employed appearance features. Although all above works perform favorably for the nearly frontal faces, they struggle when the faces are confronted with large changes of appearance from view variations and severe occlusions. This is mainly because the shape and appearance relations exhibited in faces with large view variation are rather complex for the regression function to handle. For faces with severe occlusions, the performance of the learned model becomes even worse, because the stage-wised regressor is unable to take the outliers caused by the severe occlusion into consideration.

More recently, to address the aforementioned limitations of prior work, many algorithms based on deep convolutional neural network have been proposed which benefits from the strong discriminative deep features. In the topic of human pose estimation, which aims at locating the precise pixel location of key points of a human body, is quite close to face alignment. The architecture combining the part detection network with Hourglass [27] and deep regression network has achieved huge success in human pose estimation [4], and naturally has been applied to face alignment [3]. The Hourglass Network is based on bottom-up, top-down procession along with intermediate supervision, which enables the network to capture features from differ-

Figure 1. First row, some faces with large view angles in the menpo dataset. Second row, the normalised face images by supervised transformation. Third row, the alignment results of Stacked Hourglass Model.

ent scales as well as contextual information. Therefore, this detection followed by regression method is capable of dealing with occlusion and large pose variations well, but the problem existed is that the speed is slow, and the model is large, reaching at hundreds of MBs.

In this paper, to provide a robust face alignment algorithm, we combine the supervised face transformation [9] and Stacked Hourglass Network [27]. Firstly, the supervised face transformation aims at reducing the shape variance among the datasets by removing rigid transformations from translation, scale and rotation. If the shape variance in datasets is large, as at the first row in Figure 1, the learning model will be deteriorated. After the supervised face transformation, the new training dataset is illustrated as the second row in Figure 1. Therefore, the regressors learning process will pay more attention to the feature and non-rigid transformation. Secondly, the Stacked Hourglass Network based on convolutional layers is explored in our algorithm, which serves as discriminative feature provider and strong regressor. To better evaluate our algorithm, we re-implement two cascade shape regression based algorithms LBF and SDM with supervised face transformation to make a comparison. Extensive experiments on four challenging datasets, *COFW*, *IBUG*, *300W* and the Menpo Benchmark, confirm the effectiveness of the proposed methods.

2. Related Work

The proposed method belongs to the regression based method. In this section, we mainly review two kinds of popular regression based methods, the cascade shape regression and the convolutional neural network based methods.

2.1. Cascade Shape Regression

The main idea of Cascade Shape Regression (CSR) [14, 7, 39] is to learn a sequence of regressors in an additive manner to approximate an intricate nonlinear mapping between the initial shape and the ground-truth shape. Specifically, given a set of N images $\{I_i\}_{i=1}^N$ and their corresponding ground truth $\{\widehat{\mathbf{x}}_i\}_{i=1}^N$, in which $\widehat{\mathbf{x}}_i = [x_1, y_1, ..., x_l, y_l, ...x_n, y_n]^\top$ and n is the number of the facial landmarks. A typical CSR model is formulated as:

$$\mathbf{W}^t = \arg\min_{\mathbf{W}^t} \sum_{i=1}^{N} \sum_{j=1}^{L} \|(\widehat{\mathbf{x}}_i - \mathbf{x}_{ij}^{t-1}) - \mathbf{W}^t \Phi(I_i, \mathbf{x}_{ij}^{t-1})\|_2^2, \tag{1}$$

where $\mathbf{W}^t$ is the linear regression matrix, which maps the shape-indexed features to the residual shape. $\mathbf{x}_{ij}^{t-1}$ denotes the intermediate shape of image I_i at stage $t-1$. j counts the perturbations. The training data is augmented with L multiple initialisations for each image, which serves as an effective method for improving the generation capability of training. t is the current iteration number. $\Phi(I_i, \mathbf{x}_{ij}^{t-1})$ denotes the shape-index feature. The CSR model can be solved by the least square algorithm.

In the test procedure, CSR is performed sequentially closer to the ground-truth shape as in Algorithm 1.

Algorithm 1 Cascade Shape Regression

Input: Face image: I, initial shape estimation: $\mathbf{x}^0$, shape-indexed feature extractor: Φ.

1. **for** $t = 1$ to T
2. compute shape-indexed features: $\Phi(I, \mathbf{x}^{t-1})$
3. residual shape estimation: $\Delta\mathbf{x} = \mathbf{W}^t\Phi(I, \mathbf{x}^{t-1})$
4. update update: $\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}$
5. **end**

Output: final estimation $\mathbf{x}^T$.

One critical factor of CSR is the informative shape-indexed features, which can be achieved by feature selection/learning method or off-the-shelf feature descriptor. Cao *et al*. [7] propose shape-indexed features and a correlation-based feature selection method to learn informative features. Ren *et al*. [29] utilise the random forest to learn discriminative binary features with "locality" principle, and it is worth mentioning that this method achieves 3000 fps speed. Xiong *et al*. [39] concatenate the SIFT features around each landmark as the shape-index feature and learn the regression matrix via linear regression. P. Burgos-Artizzu *et al*. [5] find that CSR is sensitive to occlusion, they present a robust cascaded pose regression method, which incorporates occlusion directly during learning to improve shape estimation so that the robust shape-indexed features can be used.

Another key point is the model capacity. Because CSR works in a data-driven manner, which means it remembers the fitting paths in the training phase and directly maps face

appearance to the target shape update in the testing phase. However, over-fitting occurs when a discrepancy exists between the fitting rates in learning and testing. Therefore, Xiong *et al*. [40] enhance the model's robustness by dividing the whole search space into individual regions with similar gradient directions. This method enhances the model capacity by learning different models in a "divide and conquer" strategy. Yang *et al*. [41] propose a deep convolutional network model to estimate head pose so that an alignment-friendly initialisation is allocated. This method enhances the algorithm's performance from a lateral view by adding the pose prior information to decrease the shape variance.

2.2. Convolutional Neural Network

To detect facial landmarks in near-frontal faces, Sun *et al*. [35] first apply a three-level convolutional network to obtain robust and accurate landmark estimation from coarse to fine. Zhang *et al*. [43] have modeled a multi-task problem to deal with the facial landmark location and attribute classification. Trigeorgis *et al*. [36] have applied the recurrent neural networks to face alignment.

For large pose face alignment problem, [19] proposes an algorithm after combining the cascaded CNN regressor method with 3DMM, in which CNN-based regressors are used to estimate camera projection matrix and 3D shape parameters while face alignment is seen as a 3DMM fitting problem. Also, Zhu *et al*. [45] present a 3D solution for this issue, in which an iterative manner through a single CNN is explored. In [3], the face alignment is treated as a two-step processing. The first facial part detection provides confidence scores of each landmarks, which is seen as a local evidence, while the following regression networks are aggregated to produce final prediction, which is a global regression. Therefore, both the part details and global context information are fully explored.

3. The Proposed Method

3.1. Overview

Figure 2 is an overview of the proposed method. It is made up of two main steps. The first step focuses on supervised face transformation, which aims at reducing the overall variations from rigid transformations of the datasets. This is much inspired by the characteristics of training datasets, we find that there are some faces with large view angles (in Figure 1), adding the difficulty of accurate alignment, mainly because alignment algorithms are sensitive to initialisation. The second step of the proposed method is a deep convolutional networks consisting of four stacked Hourglass Networks. The Hourglass Network is capable of extracting multi-scale discriminative feature in a human face, and also functions as a regressor to locate the final landmarks. The following sections will describe the above steps in detail.

3.2. Supervised Face Transformation

In our experiments, we have found that faces with large view angles will deteriorate the model during the learning period. These images are common in face alignment datasets, shown in Figure 1.

In our system, a face detector [9] trained on 400K face images is firstly used to detect the five semantic facial landmarks, simply named as **5L**. Chen *et al*. [9] incorporate the Supervised Transformer Network into the cascade convolutional neural network to deal with the large pose variations faced in the task face detection. The **5L** is used to rotate, rescale the training sample to the mean shape, in order to relieve the rigid transformation, shown in Figure 1. Specifically, we use the following formula (Eq 2) to define a similarity transformation via Procrustes Analysis:

$$\begin{bmatrix} m_x \\ m_y \end{bmatrix} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \begin{bmatrix} x^i \\ y^i \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{2}$$

According to the value of a and b, it is easy to compute the rotating angle and scale factor. The original images are rotated and rescaled based on the parameters of the transformation.

Then, we use the facial landmark detector in [2] to detect the nineteen semantic facial landmarks, simply named as **19L** to further remove the influence from rigid transformation. The UMDfaces is a large dataset, which contains 367,920 face annotations of 8501 subjects. In addition, this dataset is diverse from the aspect of head pose, occlusion, and quality. Ankan *et al*. [2] have trained a network after adapting the VGG-Face architecture, and have publicised it on https://www.umdfaces.io. In our experiment, we use this ready-made model to predict **19L**. **19L** further helps to provide alignment friendly initialisation, in which the face is copped with size (256×256), and the face is located at the centre of the image. The cropped images are fed to the following stacked Hourglass Network as training set.

3.3. Stacked Hourglass Network

Next, the state-of-the-art architecture Hourglass network proposed in [27] is employed to estimate the location of each landmark. The component in Figure 2 is a single Hourglass, as shown in Figure 3. It is a four level structure, based on the residual network [18]. The residual module is able to extract high level feature based on the convolutional operation, at the same time, it can retain the original information with the skip route. It only changes the depth of the data without changing the size of data. Therefore, it can be seen as an advanced convolution layer.

The structure of Hourglass in Figure 3 is a symmetric topology, so it is able to capture and consolidate information from different scales and resolutions. Before down-

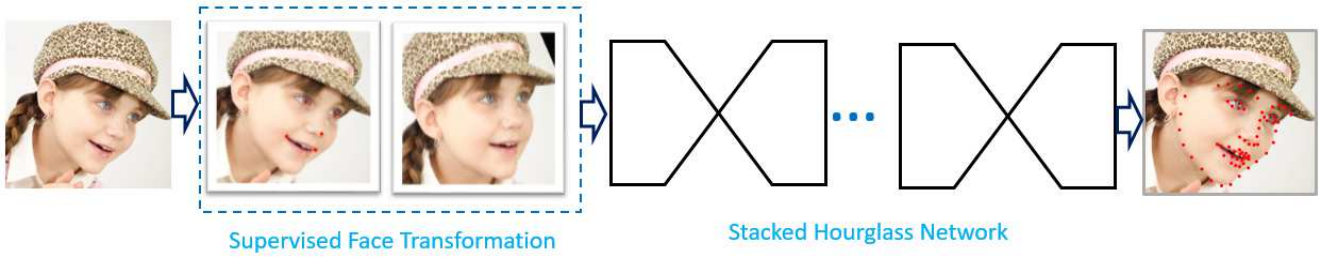


Figure 2. The proposed method in this paper includes two parts: a supervised transformation to decrease the regression target and a Stacked Hourglass Model to increase the capacity of the regression model.

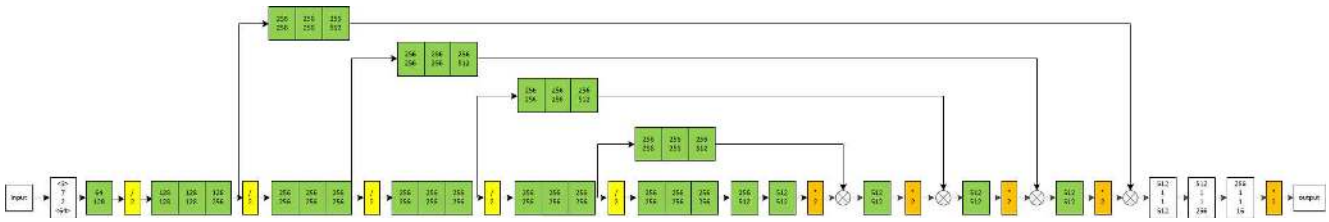


Figure 3. The structure of the Hourglass.

sampling operation, it separates a single route to retain the information in the current size. Before upsampling operation, it adds the maps with the same size from the original layer. Between the processing of two downsampling operations, it uses three residual modules. After adding the two maps, another residual module is employed to extract the feature.

These local evidence is essential for identifying each semantic landmark in a human face. The person's poses, expressions, and the relationships of the individual landmarks are better explored at different scales in an image. Also, the heatmaps cover both the coordinate and confidence information. As is discussed in [4], the confidence of the occluded parts is lower than that of the visible ones. Therefore, it is able to estimate the occluded points by taking the contextual information into consideration. We stack four Hourglass Networks in the Menpo challenge competition. After the fourth Hourglass Network, we can directly obtain the face shape. Then the inverse supervised face transformation is used to obtain face shape on original image.

Empirically, we determine that a value of 5px to be optimal for a face size of 256×256. We train our facial landmark localisation with the following L2 loss:

$$\ell_2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{jk} \|\widehat{\mathbf{x}}_i(j,k) - \mathbf{x}_i^*(j,k)\|^2, \qquad (3)$$

where $\widehat{\mathbf{x}}_i(j,k)$ and $\mathbf{x}_i^*(j,k)$ represent the predicted and the ground truth confidence maps at pixel location (j,k) for ith part, respectively.

4. Experiments

4.1. Datasets

In this paper, the training data are only from the Menpo Benchmark [30] [31], which contains 2300 profile images with 39 points and 6679 images with 68 points. The face images are in arbitrary poses, and cover both (near) frontal, as well as profile faces in the wild.

We evaluate the proposed methods on public datasets (*COFW*, *IBUG*, *300W* indoor and outdoor).

COFW [5] [16] focuses on occlusion. Commonly, there are 845 faces from *LFPW* training set and extra 500 faces with heavy occlusion in training set. The testing set is made up of 507 images that are heavily occluded. The dataset contains internet photos depicting a wide variety of more difficult poses and includes a significant amount of occlusion. We conduct evaluations on 68 points [17].

300W [33] is short for 300 Faces in-the-wild. Note that challenging subset is extremely difficult as its images have large variations in face poses, expressions and illuminations. Here, We only conduct evaluations on 68 points on challenging subset (IBUG). Besides, evaluation are also performed on the newly updated *300W indoor and outdoor* database [32] that consists of 300 Indoor and 300 Outdoor in-the-wild images. It covers a large variation of identity, expression, illumination conditions, pose, occlusion and face size.

Menpo Benchmark Testset The test data of the Menpo Benchmark includes 4253 profile faces and 12006 frontal faces. We submit the predictions by Deep_Yang and get the result from the organiser.

4.2. Baseline Methods

The supervised transformation is able to decrease the shape variances and increase the capacity of the regressors. We recomplement two CSR-based algorithms to make comparisons between handcraft features, learning features and convolutional features. The methods we recomplement are named as LBF_Yang and SDM_Yang.

LBF_Yang: We present a fast model based on JDA [10] and LBF [29]. In [29], Ren *et al*. have proved that alignment friendly face rectangle (bounding box) is beneficial to the cascade shape regression in face alignment. Based on the insight, we construct a fast face alignment system in android mobile phone based on the JDA face detector [10] and LBF face alignment model [29]. The whole model in mobile phone is about 45MB in total, consisting the face detection model (10MB), the tracking model and the detection model (30MB). It runs at about 400fps in HuaWei P10 plus. Different from the common routine that uses the JDA as face detector, we use the five landmarks predicted by JDA to produce a face rectangle estimation.

SDM_Yang: In addition, we also implement the SDM [39] algorithm with some alterations. First, we use the multi-scale SIFT feature whose dimension is 128 for each landmark. Besides, instead of using the shape residual as the regression goal, in our SDM, we employ the normalised shape residual as the regression target. That is to say, we normalise the current shape and target shape to the mean shape, then the difference between the normalised shapes are the new learning target. By doing so, the shape residual from the rigid transformation can be removed, our learning algorithm will pay more attention to the non-rigid transformations from other factors like pose variations, occlusions, and to name a few. Furthermore, the initial shape in our algorithm is from the five points provided by JDA. This algorithms runs at about 30fps in computer. The model size is approximate 20MB.

4.3. Evaluation Metric

Fitting performance is usually assessed by the normalized mean error. In particular, the average Euclidean point-to-point error normalized distance is used. The normalized mean error over all landmarks,

$$E_i = \frac{\frac{1}{n}\sum_{j=1}^{n} |\mathbf{x}_{i,j} - \widehat{\mathbf{x}}_{i,j}|_2}{|\mathbf{lt}_i - \mathbf{rb}_i|_2}, \tag{4}$$

where n is the number of landmarks, $\mathbf{x}_{i,j}$ is the prediction, $\widehat{\mathbf{x}}_{i,j}$ is the ground truth. $\mathbf{lt}$ and $\mathbf{rb}$ are the positions of the left top point and right bottom point of the bounding box of ground-truth shape. The normalization is able to make the performance measure independent of the actual face size or the camera zoom factor.

The cumulative distribution function (CDF) of the normalized root mean squared error (NMSE) is employed for performance evaluation.

4.4. Evaluations on Public Benchmarks

Figure 4(a), 4(b), 4(c), 4(d) show the quantity evaluation on the public face alignment datasets. Table 1 records the failure rate, which computes the percentage of images whose NMSE is larger than 0.08, and the mean error, which represents the performance of the algorithms to a certain extent. On *COFW*, NMEs of Deep_Yang, SDM_Yang, LBF_Yang are 1.8%, 2.2%, 2.6% respectively. Deep_Yang ranks 1st, because it benefits from the heatmap which instructs the whole network to pay more attention to the feature with high confidence in an explicit manner, increasing its robustness to occlusion. On the other three datasets, Deep_Yang obtains the best performance, which shows that deep feature increases the model capacity when faced with large pose variations.

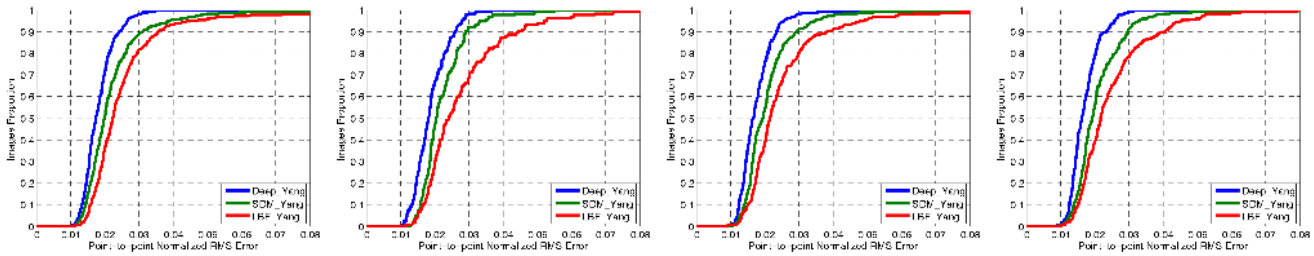
Overall, We find that the deep model ranks 1st in all above datasets. This illustrates that the deep features from the convolutional layers are more informative. The discriminative multi-scale features extracted from the Hourglass Networks, which are able to cover the important information needed in landmark localisation. Besides, as is listed in Table 1, the overall failure rate is low because the supervised face transformation based on efficient face detector boosts algorithms' capacity.

Although the CSR-based methods do not achieve superior performance, they still gain a comparatively excellent performance. They have advantages over deep model from at least two aspects. On the one hand, model size of SDM can be compressed, [23] has compressed the model to several MBs. By comparison, the deep model is large in size about hundreds of MBs. On the other hand, LBF_Yang, which runs at thousands frames per second in computer and also can run hundreds frames per second on mobile phone. In comparison, when deep model is used to predict facial landmarks, it takes few seconds per frame.

4.5. Menpo Benchmark Competition

Our deep model has been evaluated independently by the organizers using their own ground truth which are not disclosed to the participants. Figure 5(a) and 5(b) illustrate that our deep model yields better performance in both the semifrontal and profile testing datasets. Also, our algorithm can be applied to the facial landmark tracking system as in [42].

Besides, we also show some examples on the testing datasets in the Menpo Benchmark in Figure 6(a) and Figure 6(b). We find that our deep model can deal with the challenges like illumination, pose, occlusion well.

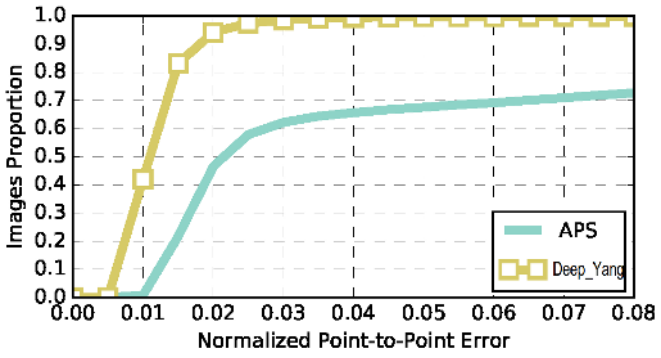


(a) COFW (b) IBUG (c) 300W indoor (d) 300W outdoor

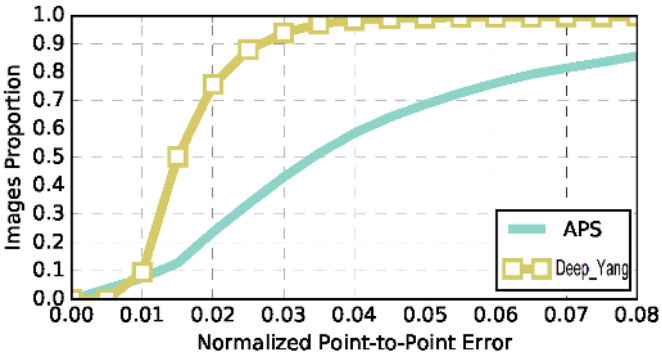
Figure 4. Validation experiments on COFW, IBUG, 300W indoor and outdoor data sets.

	COFW		IBUG		300W Indoor		300W Outdoor	
Method	NME(%)	FR (%)	NME(%)	FR (%)	NME(%)	FR (%)	NME(%)	FR (%)
Deep_Yang	1.8/4.0/5.6	0	1.9/4.9/7.0	0	1.8/4.1/6.1	0	1.7/4.0/5.8	0
SDM_Yang	2.2/4.8/6.8	0.4	2.2/5.7/8.2	0	2.1/4.9/7.2	0	2.1/5.0/7.2	0
LBF_Yang	2.6/5.7/8.0	1.8	2.8/7.2/10.4	0.7	2.5/5.9/8.6	1	2.5/5.9/8.5	0

Table 1. Landmark localisation results on four public test sets using 68 points. Accuracy is reported as the Normalised Mean Error (NME) and the Failure Rate (FR). To facilitate comparison with other methods on these datasets, we give mean error normalised by the diagonal of the ground truth bounding box, the out eye corner distance and the eye centre distance.



(a) Semifrontal



(b) Profile

Figure 5. Evaluation on the test sets of Menpo Benchmark.

5. Conclusion

We have performed an experimental comparison of the proposed three methods based on two popular frameworks for face alignment. We hope this will help practitioners choose an appropriate method when deploying face alignment system in the real world. We have also present a technique for improving face alignment algorithms by supervised face transformation.

Acknowledgments We would like to thank Dong Chen from Microsoft Research Asia for providing face detection results on the Menpo challenge dataset. The work is supported in part by the Natural Science Foundation of China under Grant 61532009 and 61402233, in part by the National Science Foundation of Jiangsu Province under Grant BK20151529.

References

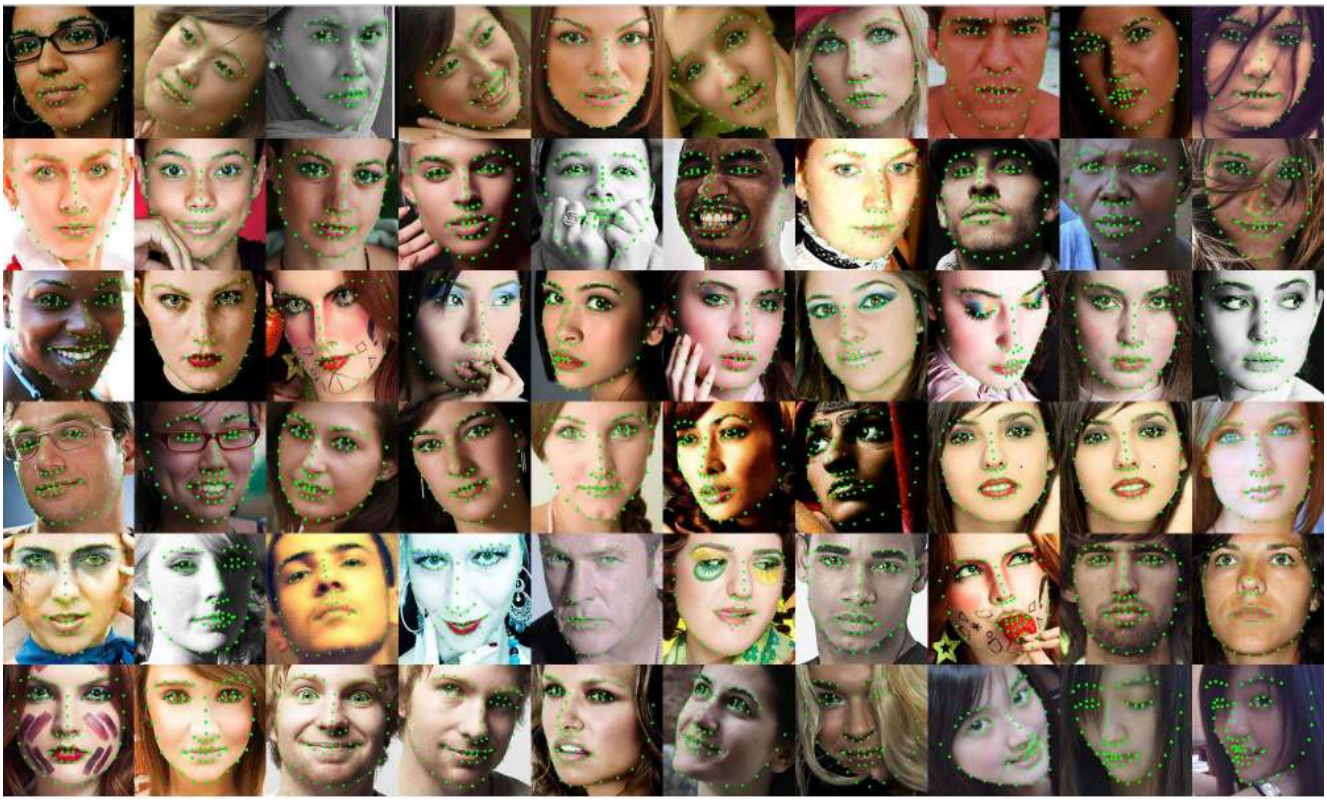
[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866. IEEE, 2014. 1

[2] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. *arXiv preprint arXiv:1611.01484*, 2016. 3

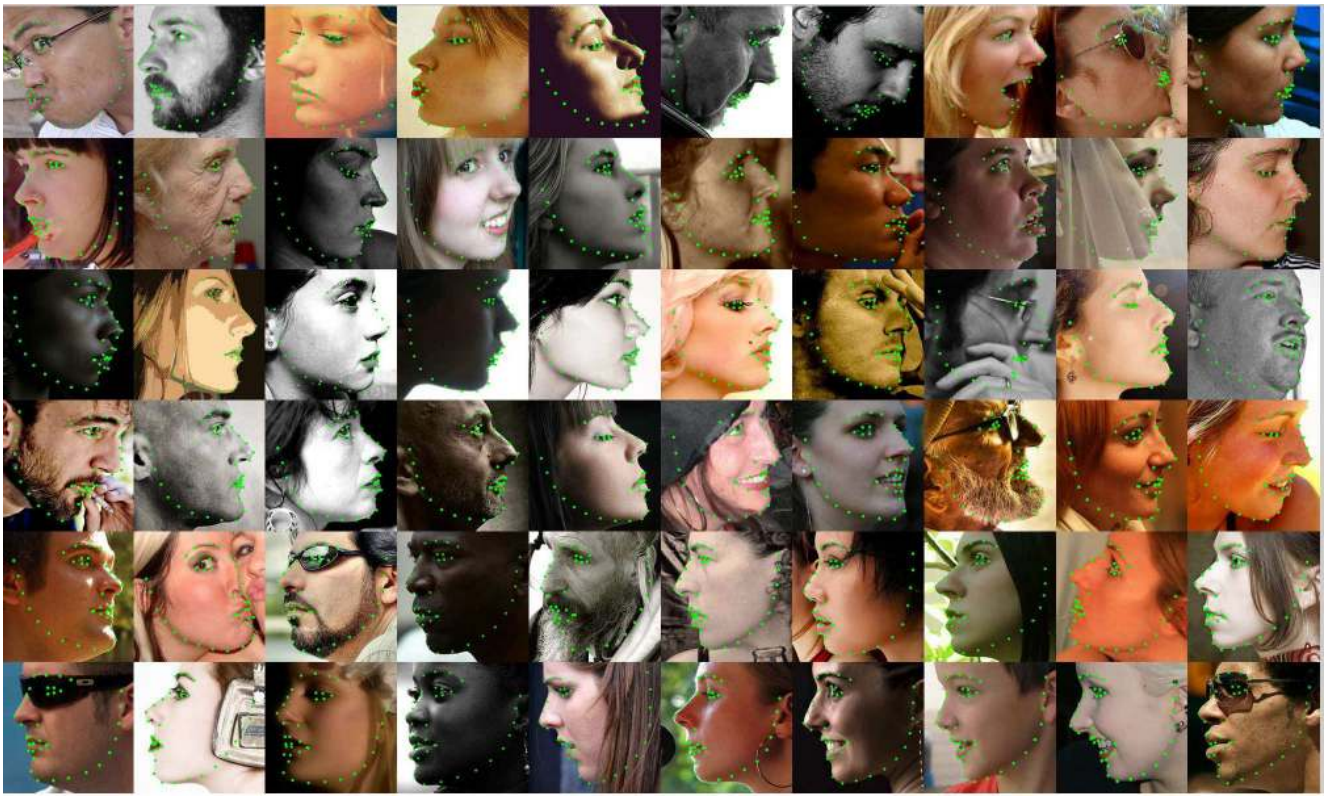
[3] A. Bulat and G. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *Proceedings of British Machine Vision Conference*, 2016. 1, 3

[4] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 1, 4

[5] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520. IEEE, 2013. 2, 4



(a) Semifrontal



(b) Profile

Figure 6. Example results on the test sets of Menpo Benchmark.

[6] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics (TOG)*, 35(4):126, 2016. 1

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894. IEEE, 2012. 1, 2

[8] O. Çeliktutan, S. Ulukaya, and B. Sankur. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13, 2013. 1

[9] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016. 2, 3

[10] 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. 5

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):681–685, 2001. 1

[12] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *Proceedings of European Conference on Computer Vision*, pages 278–291. Springer, 2012. 1

[13] J. Deng, Q. Liu, J. Yang, and D. Tao. M 3 csr: multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47:19–26, 2016. 1

[14] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1078–1085. IEEE, 2010. 1, 2

[15] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 300–305, 1998. 1

[16] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2385–2392, 2014. 4

[17] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 4

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of theComputer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[19] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, 2016. 3

[20] V. Kazemi and J. Sullivan. Face alignment with part-based modeling. In *Proceedings of British Machine Vision Conference*, pages 27–1. British Machine Vision Association, BMVA, 2011. 1

[21] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 365–372, 2009. 1

[22] L. Liu, C. Xiong, H. Zhang, Z. Niu, M. Wang, and S. Yan. Deep aging face verification with large gaps. *IEEE Transactions on Multimedia*, 18(1):64–75, 2016. 1

[23] Q. Liu, J. Deng, and D. Tao. Dual sparse constrained cascade regression for robust face alignment. *IEEE Transactions on Image Processing*, 25(2):700–712, 2016. 5

[24] Q. Liu, J. Yang, J. Deng, and K. Zhang. Robust facial landmark tracking via cascade regression. *Pattern Recognition*, 66:53–62, 2017. 1

[25] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2864–2871, 2013. 1

[26] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004. 1

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 1, 2, 3

[28] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445, 2000. 1

[29] 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*, pages 1685–1692. IEEE, 2014. 1, 2, 5

[30] e. a. S. Zafeiriou. The menpo facial landmark localisation challenge. In *Computer Vision and Pattern Recognition Workshops(CVPRW), 2017*. IEEE. 4

[31] G. C. J. D. S. Zafeiriou, G. Trigeorgis and J. Shen. The menpo facial landmark localisation challenge: A step closer to the solution. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017. 4

[32] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 4

[33] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 4

[34] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014. 1

[35] 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*, pages 3476–3483. IEEE, 2013. 3

[36] 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. 3

[37] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 34(2):372–386, 2012. 1

[38] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu. Robust face recognition via adaptive sparse representation. *IEEE Transactions on Cybernetics*, 44(12):2368–2378, 2014. 1

[39] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539. IEEE, 2013. 1, 2, 5

[40] X. Xiong and F. D. la Torre. Global supervised descent method. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015. 3

[41] 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. 3

[42] J. Yang, J. Deng, K. Zhang, and Q. Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 41–49, 2015. 5

[43] Z. Zhang, P. Luo, L. C. Change, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014. 3

[44] X. Zhao, X. Chai, and S. Shan. Joint face alignment: Rescue bad alignments with good ones by regularized re-fitting. In *Proceedings of European Conference on Computer Vision*, pages 616–630. Springer, 2012. 1

[45] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 3

[46] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 113–120, 2013. 1