3-D Face Landmark Detection using Multitask Hourglass Model

Martin Pernus, Vitomir Struc
Univerza v Ljubljani, Fakulteta za elektrotehniko, Trzaska cesta 25, 1000 Ljubljana
martin.pernus1@gmail.com

Abstract

Even though state-of-the-art 2-D face alignment methods have shown near saturating performance, solving face alignment based on 3-D structure of face from 2-D image remains challenging problem. 3-D face landmarks, which correspond with 2-D projection of face structure have advantage over 2-D landmarks of preserving correspondences across poses. In order to try to solve this task more effectively, a multi-task Hourglass model is proposed, which combines the state-of-the-art convolutional neural network model Hourglass with multi-task learning in form of learning facial landmarks and face pose simultaneously, where the pose is represented by Euler angles. Experiments on dataset AFLW2000-3D show that the results of multi-task Hourglass improve the results compared to single-task Hourglass, learned on landmarks only.

1. Introduction

Facial landmark localization, also known as face alignment is arguably one of the most heavily researched topics recently. Given a single image, the face alignment algorithm will try to determine the pixel locations of several face points. This task is beneficial to many higher level tasks, for example attribute analysis [21], expression analysis [24] and face recognition [19, 29, 30].

Facial landmark localization has performed impressively in tasks, regarding to localization of 2-D landmarks, which correspond to facial points, observed from an image. However, the spatial distribution of the 2-D facial landmarks is highly pose dependent and does not correspond to 3-D structure of the human face. In contrast, 3-D facial landmarks, which correspond to 2-D projections of 3-D facial structure, preserve correspondences across poses. In this paper, we will refer to the 2-D projections of the 3-D landmarks in the image plane as 3DA-2D landmarks to distinguish them from the 3-D coordinates of the facial landmarks in the 3-D coordinate system.

The main solutions have tried to do face alignment via 3-D morphable models [15, 16, 42] and Convolutional Neural Networks (CNNs) [3, 1]. Although 3-D morphable models can cover arbitrary poses, they are bounded by the linear parametric 3-D model, which can hurt their performance. CNNs have shown better performance, which is obtained by using a specific CNN model Hourglass in which features are processed across all scales. However, all CNN methods so far have only tried predicting landmarks based on images alone, rather than also trying to incorporate some additional knowledge.

Figure 1: Example result of landmark detection for challenging face image. The face image is fed into trained Hourglass model, which predicts facial landmarks that correspond to 3-D structure of face as well as face pose.

Multi-task learning has been proved to improve performance of CNNs. Applying external knowledge for 2-D face landmarks detection has been successfully applied in [28, 38, 40, 39], improving results comparing to methods, that only learn single task. By applying similar methods to 3-D face alignment, performance is expected to increase.

Exploratory work is needed to find out, whether multi-task learning can further push bounds of 3DA-2D face alignment. Such multi-task learning could potentially achieve state-of-the-art results on detection of 3DA-2D landmarks. This would in turn provide better performance of higher level tasks.

In this paper, we try to address the gap between multi-task learning for CNNs and Hourglass model for prediction of 3DA-2D facial landmarks. First, the knowledge of pose orientation is extracted. This is done by considering available information about landmark locations in 2-D coordinate system (in image plane), 3-D landmark locations

and by estimated intrinsic parameters of the camera. This gives us enough information to extract pose of face in compact form of Euler angles. Then, the Hourglass model is constructed and jointly trained on 3DA-2DA landmark locations and Euler angles.

The main contributions of this paper are:

- We incorporate the knowledge about face pose into landmark prediction.
- We optimize the Hourglass model for prediction of facial landmarks using multi-task learning.

The rest of the paper is structured as follows: in section 2, we review the history of face alignment methods, origin of Hourglass model and multi-task learning. In section 3, we introduce the terminology and methods used throughout this paper. Section 4 presents the experimental setup and results. Finally, section 5 presents the open issues and summary of our work.

2. Related Work

Here, related works of our method are presented. Specifically, we focus on history of face alignment methods, CNN's specific Hourglass model and multi-task CNNs.

2.1. 2-D Face Alignment

A large number of approaches have been proposed to tackle with the problem of face alignment. We describe some of the commonly used approaches.

Most well known approach is active appearance model, which was first proposed by Cootes et al.[7]. They are linear statistical model of both the shape and the appearance of the deformable object. [22] made extensions to active appearance model and used it to locate features in frontal views of faces. [26] utilized nonlinear active appearance model, which improved its performance significantly. Main drawback of active appearance models is their troubles with partial occlusions.

Cascaded regression is another very popular method for face alignment due to its high accuracy and speed. Regression process is divided into stages by learning a cascade of vectorial regressors. Shape-indexes features, which depend on previous shape estimate can be designed by hand [31] or can be learned [5].

Ensemble regression-voting is a method that jointly estimates the whole face shape from images, during which the shape constraint is implicitly exploited. Votes are cast for the face shape from image patches via regression. Robust prediction is obtained by votes from different regions. In [8], random forest regression is used to generate feature response images, which are then used to fit a shape model. [9] further improved performance by using conditional regression forests, which are conditioned based on global face properties. [33] contributes by adding two types of sieves for filtering out votes. First type filters out votes that don't agree with hypothesis for the location of the face center, while the second one filters out distant votes.

Sun et al. [27] proposed facial alignment method with CNN to predict five face keypoints . Specifically, three-level deep CNN was used, with the first level being used to make predictions on whole face, while the remaining two levels refined the initial estimation of keypoints. [41] used similar structure to predict 68 points on face, while also improving the performance with design of coarse-to-fine network cascade and geometric refinements. In [37], several stacked auto-encoder networks are used to predict keypoints, each refining output of previous network by using multi-resolution approach.

For thorough survey of aforementioned and many other 2-D face alignment methods, reader is referred to [14].

2.2. Hourglass Model

Alejandro et al. [23] proposed novel CNN architecture called hourglass model for human pose estimation, achieving state-of-the-art results. [3] employed Hourglass model for face alignment task. In [34], the results of Hourglass model are improved by using face transformation in order to reduce shape variance of faces. Hourglass was also used in [10], which estimated both semi-frontal and profile facial landmarks, capitalising on the correspondences between the profile and frontal facial shapes.

2.3. Multi-task Convolutional Networks

Multitask learning was first analyzed by Carauna [6]. Since then, several approaches in Computer Vision used it for solving various tasks.

Zhang et al. [39] proposed deep multi-task learning for detection of facial landmarks . Optimization of facial landmark detection was performed together with head pose estimation, gender classification, age estimation and other tasks. Such learning outperformed other methods and achieved state-of-the-art results. [38] boosted performance of multi-task network by acknowledging the inherent correlation between face detection and face alignment and novel online hard sample mining strategy, which enabled real-time performance. [28] was inspired by [39], but used more landmark points while also improving performance by using network's output as robust initialization, using them to perform several iterations with network. Authors of [39] refined their existing multi-task learning method in [40] by incorporating dynamic task coefficients and new objective function, which drastically reduced model complexity, while also boosting performance.

In [25], an all-in-one CNN is presented, solving the tasks of face detection, landmark localization, pose estimation, gender recognition, smile detection, age estimation and face

verification and recognition.

2.4. 3-D Face Alignment

In order to advance face alignment in face images with arbitrary poses, estimation 3DA-2D face landmarks was developed. First methods used 3-D morphable models [15] [16], [42], which were fitted to a 2D image. [15] estimated both 2D and 3DA-2D landmarks. After integrating 3-D deformable model, cascaded regressor approach was designed to estimate the camera projection matrix and the 3DA-2D landmarks. [16] extends [15] by fitting a dense 3-D morphable model, employing CNN as the regressor, using 3DA-2D-enabled features and estimating cheek landmarks. [42] proposed fitting a dense 3-D face model to the image via CNN. A method of synthesizing large-scale training samples in profile views was also proposed, with aim of solving labelling landmarks in large poses.

3DA-2D alignment methods are bounded by parametric 3-D model, and the invisible landmarks are predicted based on 3-D morphable model fitting results on the visible appearance. Contrary to that, Bulat et al. [2] directly utilized stacked Hourglass model to predict 3DA-2D facial landmarks, achieving state-of-the-art results. In [1], the performance was improved by first rotating faces into upright position, then feeding images into two-stacked Hourglass model, where the first Hourglass estimated the 2-D landmarks, while the second one predicted the final 3DA-2D landmarks.

3. Proposed Method

3.1. Overview

Figure 2 is an overview of the proposed method. The image is an input in CNN Hourglass model, which is capable of extracting multi-scale discriminative feature in a human face due to convolution operations on multiple resolutions. The Hourglass model is trained on landmarks as well as Euler angles, which is expected to increase its performance. The landmarks are trained in form of heatmap regression. Each heatmap then gets averaged and connected to fully-connected multilayer perceptron with Euler angles as output. The heatmaps that belong to invisible landmarks should produce lower average value than the heatmaps of visible landmarks, which the fully connected multilayer perceptron could capitalize on.

In order to get pose information in form of Euler angles, the Perspective-n-Point problem is first solved to obtain the projection matrix. The projection matrix has three degrees of freedom. It can be most compactly represented with three-dimensional vector of Euler angles.

The following sections will describe the above steps in detail.

3.2. Pose Estimation

In estimating the pose, the problem boils down to finding the pose of an object when we know the intrinsic camera parameters, 3-D locations of landmarks in arbitrary reference coordinate system (world coordinate system) and 2-D locations of those landmarks in image plane. The intrinsic camera parameters consist of effective focal lengths in pixels f_x and f_y, location of principal point expressed with x_0 and y_0, and skew γ. Together they form the camera parameters matrix:

$$\boldsymbol{K} = \begin{bmatrix} f_x & \gamma & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

The problem of determining the pose of an object is often referred to as Perspective-n-Point (PnP) problem and can be solved using variety of algorithms. We chose to use perspective-three-point algorithm [11], which gives us required orientation of camera in the world coordinate system. The orientation is generally given in form of 3x3 rotation matrix, which has 3 degrees of freedom and can be represented using 3-dimensional Euler angles. Given 3x3 rotation matrix

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{2}$$

one can compute Euler angles as:

$$\theta_x = \text{atan}_2(r_{32}, r_{33}), \tag{3}$$

$$\theta_y = \text{atan}_2(-r_{31}, \sqrt{r_{32}^2 + r_{33}^2}), \tag{4}$$

$$\theta_z = \text{atan}_2(r_{21}, r_{11}), \tag{5}$$

where atan_2 denotes the quadrant checking arc tangent function.

3.3. Hourglass Model

Next, the state-of-the-art architecture Hourglass, proposed in [34], is constructed. The Hourglass model is a symmetric CNN that is able to capture and consolidate information from different scales and resolutions. Our model consists of four downsampling and four upsampling operations. Before each down-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. The fundamental block of Hourglass structure is called residual unit [12], as shown in figure 3. Each residual unit acts as a small neural network with a skip connection, thanks to which the signals can make its way across the whole network easier. The residual unit is employed after downscaling, upscaling and on each separate route.

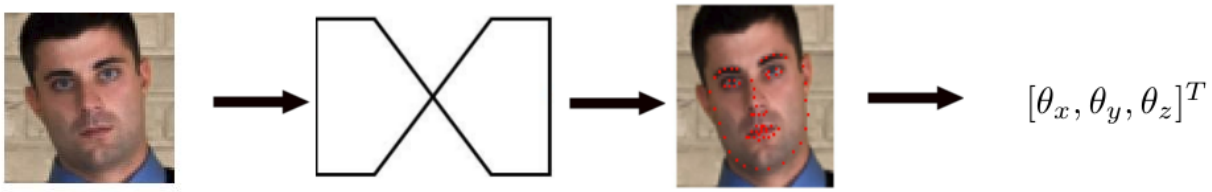

Figure 2: The proposed method in this paper. The input image gets fed into Hourglass model, which will predict facial landmarks as well as pose in form of Euler angles. Euler angles are estimated based on predicted heatmaps for each landmark.

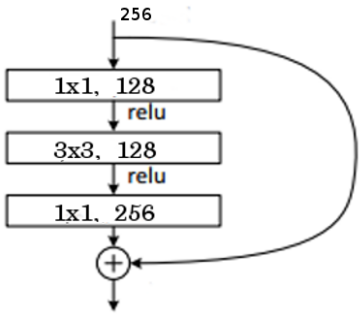

Figure 3: Residual unit, used as fundamental block in the Hourglass model. It is characterized by skip connection, thanks to which the signals can make its way across the whole network easier. In comparison with [12], our implementation of first 1x1 and 3x3 convolutions consist of 128 feature maps instead of 64.

After the Hourglass structure, the landmarks are predicted in form of heatmap regression, where each landmark is represented as two-dimensional Gaussian probability density function, where the maximum is at the pixel corresponding to the current landmark. 2-D global average pooling [20] is employed on predicted heatmaps and connected to multilayer perceptron, which predicts the three-dimensional Euler angles.

The loss function for landmarks is defined as L2 loss as:

$$J_1 = \frac{1}{N_L} \sum_{i=1}^{N_L} \sum_{jk} ||x_i(j,k) - \hat{x}_i(j,k)||^2, \qquad (6)$$

where N_L represents the number of landmarks, $x_i(j,k)$ and $\hat{x}_i(j,k)$ represent the ground truth confidence map in form of 2-D Gaussian probability distribution and predicted confidence map respectively at pixel location (j,k) for i-th landmark. The loss function for Euler angles is defined as:

$$J_2 = \frac{1}{3}||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||^2, \qquad (7)$$

with $\boldsymbol{\theta} = [\theta_x, \theta_y, \theta_z]^T$ and $\hat{\boldsymbol{\theta}} = [\hat{\theta}_x, \hat{\theta}_y, \hat{\theta}_z]^T$ as vectors of ground truth Euler angles and estimated Euler angles respectively.

The total loss J, which is optimized by multi-task learning is defined by

$$J = J_1 + \kappa J_2, \qquad (8)$$

where κ denotes the importance of optimizing task of optimizing Euler angles with respect to the task of optimizing landmark locations.

4. Experimental Results

4.1. Datasets and Performance Metric

In order to have comparable results, the training and testing dataset were identical as the ones in [1], which currently has the state-of-the-art results.

For training, Menpo Benchmark dataset [36] is used, which consists of 5658 semi-frontal and 1906 profile facial images. The images were selected from FDDB [13] and AFLW [18] datasets. The annotated face images are collected from unconstrained conditions, which exhibit large variations in pose, expression, illumination, etc. The Menpo Benchmark dataset is annotated with 68 and 39 2-D landmarks for semi-frontal and profile faces respectively. In order to evaluate our method in 3DA-2D setting, the 68-landmark 3DA-2D annotation scheme from [2] is used.

The Menpo Benchmark dataset was also annotated in [35], where 84 landmarks were represented as 3DA-2D points on image coordinate system as well as 3-D points in model space coordinate system. This annotation scheme was used to estimate the pose of each face.

For testing dataset we used the AFLW2000-3D dataset [42]. It contains 2000 face images captured in the wild with large pose variations, severe occlusions and extreme illuminations, annotated with 68 3DA-2D landmarks.

As a performance metric, the Normalized Mean Error (NME) is utilized [42, 15]. Whereas the landmarks position error used to be normalized by the distance between the eyes, the metric NME was introduced to normalize faces where the distance between eyes is not present due to ex-

treme face poses. It is defined as:

$$\text{NME} = \frac{1}{N_T}\sum_{m=1}^{N_T}\frac{1}{d_k}\frac{1}{N_L}\sum_{i=1}^{N_L}||[u,v]^T - [\hat{u},\hat{v}]^T||, \quad (9)$$

where N_T is the number of testing samples, d_k is the square root of the face bounding box area for the k-th testing sample, (u_{jk}, v_{jk}) and $(\hat{u}, \hat{v})$ are, respectively, the ground truth and estimated coordinates for i-th landmark.

4.2. Experimental Setup and Results

4.2.1 Pose Estimation

The annotation scheme of [35] provides us data about 84 facial landmarks in 3DA-2D image coordinate system as well as 3-D model space coordinate system. To calculate the 3-D pose of a face, we still require the information about intrinsic parameters of the camera. Since those are not known, we approximate the optical center $[x_0, y_0]$ by the center of the image, approximate the focal lengths $[f_x, f_y]$ by the width and height of an image and set skew γ to 0. Note that approximation of focal lengths in such way is optimistic at best, however required to get an estimation of pose. The orientation of camera was obtained using MATLAB's perspective-three-point algorithm in form of 3x3 rotation matrix, which was converted into Euler angles, measured in radians.

4.2.2 Training Setting

A face detection algorithm [38] was employed to obtain bounding box of the face. The bounding box is extended by factor 1.3 to capture all facial landmarks due to extreme poses. The face detection algorithm detected 89 % of faces in the training dataset. The images, where face detector failed to find a face, were cropped to square shape. The images were then scaled to 256×256. The network first begins with a 7×7 convolutional layer with stride 2 and zero padding, followed by leaky rectified linear unit activation function [32] and max-pooling with stride 2 to bring the resolution down from 256 to 64. We observed that such resolution reduction reduces GPU memory usage, allowing us to use larger batch size and reducing the size of the model.

The last convolutional layer of Hourglass model is connected to 3×3 convolutional layer, which acts as an output for landmark locations in form of heatmap for each landmark. The heatmap is constructed as discrete two-dimensional Gaussian probability function with variance 5 pixels and maximum in the pixel, that corresponds to current landmark. The heatmap is scaled in such way that its maximum equals 1. Global average pooling is then employed to extract information from heatmaps and connect it to multilayer perceptron with two densely connected layers. The first layer consists of 50 neurons with rectified linear unit as activation function. The second layer uses linear activation function and acts as the output for the Euler angles.

	Average pooling	Maximum pooling
$\kappa = 0$	7.45	7.45
$\kappa = 10^{-1}$	5.56	5.37
$\kappa = 10^{-2}$	5.45	5.51

Table 1: Results of our model in NME (%), comparing different coefficients κ and methods of pooling heatmap layer.

The model was trained using Keras with TensorFlow backend, batch size of 32 and 70k learning steps. Adam's stochastic optimization algorithm [17] was found to be the best for our task. Initial learning rate is 10^{-4} and drops to 10^{-5} after 50 epochs. A mean squared error loss is applied to compare the predicted heatmaps and Euler angles to ground-truth ones. A learning step takes approximately 0.7 seconds on one NVIDIA GeForce Titan X. During testing, face regions are again cropped using face detection algorithm, where it found 92 % of faces, and resized to 256×256. The whole training procedure of 70k learning steps takes around 14 hours. In case the total error decreased for less than 10^{-5} in the last 10 epochs, the optimization procedure stopped. It takes 7 milliseconds to generate the response heatmaps and Euler angles for a single image.

4.2.3 Results

The model was constructed and tested with three different κ coefficients, $\kappa = \{0, 10^{-2}, 10^{-1}\}$. The higher the κ coefficient, the more importance does correct pose have on the optimization procedure. In addition, besides experimenting the κ value, the global average pooling of heatmaps convolutional layer was experimentally changed to global maximum pooling. The results for landmark detection in form of NME percentages are shown in table 1, where the landmarks were detected by finding maximum value in each heatmap and then rescaled back into original image. Note that when κ equals 0, the choice between global average pooling and global maximum pooling becomes irrelevant, since that part of the network does not contribute to total loss function J. The average J_2 error on training set for positive coefficients κ decreased approximately from 1.05 to 0.80, measured in radians.

Additionally, the model was also tested when presented with augmented training data and $\kappa = 0$ with the hopes of making the model more robust. The training data was, along with ground truth heatmaps, randomly translated in interval [-20, 20] pixels, rotated for [-15, 15] degrees, scaled between [0.85, 1.15] of the original image size and randomly flipped. The results, however, show a NME of 28.96, which

Method	NME (%)
RCPR [4]	4.26
ESR [5]	4.60
SDM [31]	3.67
3DDFA+SDM [42]	3.43
Bulat et al. [2]	2.47
CMHM [1]	2.36
Ours	5.37

Table 2: Results of other methods and our method in form of NME (%).

shows severe degradation of model performance when comparing to no augmentation, $\kappa = 0$ results.

Some examples of the best and the worst performances for the testing dataset are shown in figures 4 and 5 respectively. The good performance is mostly attributed to clear images without occlusions and no large poses. The algorithm gives bad performance mostly on images with large poses and severe occlusions. Our best result is compared with the results of other techniques in table 2.

The results show that, even though the training pose error decreased relatively little on training data, it still contributes to the performance of the model. On the other hand, the differences between different pooling methods and κ values are small enough to be attributed to random variability. Our model achieves decent performance, but still needs more refinement to be comparable to state-of-the-art techniques.

5. Conclusion

In this paper, we proposed a multi-task Hourglass model for 3-D face alignment. The face pose was obtained using landmark location in 3-D coordinate system, 2-D coordinate system and estimated internal camera parameters. Face pose was compactly represented with Euler angles. The facial landmarks were trained in form of heatmap regression. Global pooling were applied to heatmaps, which was extended by multilayer perceptron with Euler angles as output. The Hourglass model was then jointly optimized on both landmarks and Euler angles. The results prove the effectiveness of multi-task learning in such configuration.

The open issues in our method evolve getting a more precise numeric representation of pose, exploring different types of blocks and structure of the Hourglass model, and improving localization by increasing the resolution of Hourglass structure. The multi-task learning could also be extended to other tasks, that could further improve the performance of the model. These topics should be addressed in future works.

References

[1] Anonymous. Cascade multi-view hourglass model for robust 3d face alignment. In *Proceedings on the IEEE International Conference on Automatic Face and Gesture Recognition*, pages X–X. IEEE, 2018.

[2] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). *CoRR*, abs/1703.07332, 2017.

[3] A. Bulat and Y. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In E. R. H. Richard C. Wilson and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 86.1–86.12. BMVA Press, September 2016.

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

[5] 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.

[6] R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[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] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. *Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting*, pages 278–291. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[9] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2578–2585. IEEE, 2012.

[10] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *CoRR*, abs/1708.06023, 2017.

[11] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):930–943, 2003.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[14] X. Jin and X. Tan. Face alignment in-the-wild: a survey. *arXiv preprint arXiv:1608.04188*, 2016.

[15] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015.

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

Figure 4: Some of the images from the testing set with the lowest average landmark error. The images lack large poses, which is the main reason for good performance of our model.

Figure 5: Some of the images from the testing set with the highest average landmark error. The reason for bad performance is in most cases due to extreme poses or face occlusions.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[19] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[20] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[21] 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.

[22] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *European conference on computer vision*, pages 504–513. Springer, 2008.

[23] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *CoRR*, abs/1603.06937, 2016.

[24] 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.

[25] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. *arXiv preprint arXiv:1611.00851*, 2016.

[26] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[27] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[28] Y. Sun, X. Zhang, and C. Li. Multi-task convolution network for face alignment. In *Journal of Physics: Conference Series*, volume 887, page 012079. IOP Publishing, 2017.

[29] 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.

[30] 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.

[31] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[32] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[33] H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[34] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *Computer Vision*

and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, pages 2025–2033. IEEE, 2017.

[35] S. Zafeiriou, G. G. Chrysos, A. Roussos, E. Ververas, J. Deng, and G. Trigeorgis. The 3d menpo facial landmark tracking challenge. In *ICCV 3D Menpo Facial Landmark Tracking Challenge Workshop*, volume 5, 2017.

[36] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *in Proc. IEEE Conf Comput. Vision Pattern Recognit. Workshops*, pages 2116–2125, 2017.

[37] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[38] 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.

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

[40] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016.

[41] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

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