Facial Landmark Detection Experiment

Yanzhe Su
University of California San Diego

Abstract: Facial landmark detection or face alignment is an important problem in computer vision. Different types of methods have been proposed to target this problem, with their advantages and disadvantages. This paper classified the past methods into three categories, briefly compared them, then attempted to reimplementing part of a recent work. Experiments were done with a common, current dataset to compare the method with other methods.

1. Introduction
Facial landmark detection or face alignment is an important problem in computer vision, and lies as a foundation step to other face-related problems like face identification and verification [1]. Given a single image containing a face, the algorithm for this problem tries to predict the precise pixel locations \(x_1, y_1, x_2, y_2, ..., x_n, y_n\) of \(n\) pre-defined feature points on the image. This problem is difficult, especially when the face is in an extreme pose, partially occluded, with special expressions or poor lightings [2].

In the past literature, different methods of facial landmark detection methods has been proposed. At first, template-based methods are used for this task [3] [4] [5]. These methods tries to learn a template from the training set to locate the keypoints, but their performance are also restricted by such templates. Later on, coordinate regression-based methods [6] [2] [7] [8] regress coordinate of the facial landmarks diretcly, in a cascaded way or alongside other tasks. Some of the methods is heavier on feature engineering, while others used deep convolutional neural networks. More recently, heatmap-based methods [9] [10] [11] are used for this task as well following the success on other image-based tasks. For the datasets used, they are also becoming more challenging with time, most notably the IBUG [12] dataset with a combination of extreme conditions, and COFW [13] dataset focused heavily on occluded faces. However, the labeling formats are inconsistent between different datasets like the two sets mentioned above, with different number of points and different meaning of each point, posed a problem on cross-dataset evaluation.

This paper made an attempt to reimplement the baseline of [14], made experiments, then compared the results with other methods. There are still accuracy gaps between the proposed method and some other recent methods. However, the other methods’ accuracy might also be questionable. In more details, Section 2 will briefly discuss some previous works about this task, in three types. Section 3 will discuss the methodology of the proposed method. Section 4 will describe the dataset used, settings and configurations of experiments done, with detailed comparison to the part this work is trying to reimplement. Finally, Section 5 will conclude the paper.

2. Related Works
Facial landmark detection has been actively researched for several decades. Before neural networks are
utilized, template-based methods like Active Shape Model (ASM) [3] [15] [16], Active Appearance Model (AAM) [4] [17] [18] [19] [20] and Constrained Local Model [5] [21] [22] are used. However, this type of methods hardly generalize to complex conditions or faces taken in the wild. More recently, with deep convolutional neural networks (DCNNs), coordinate regression-based methods and heatmap-based methods achieve much higher performance on this task.

Coordinate regression-based methods regress the location of facial landmark points directly. [2] used a course-to-fine cascaded multi-stage structure to regress the location, refining the outputs of the previous layer using networks in latter stage. [14] also used a cascaded structure, but take inputs to the local refinement networks from the feature maps in intermediate layers of its global regression network. [6] jointly optimize facial landmark detection with other tasks, predicting the landmark locations and learning other features like gender and expression simultaneously.

Heatmap-based methods predicts various properties with higher-resolution heatmap information, most commonly with a fully-convolutional structure. This type of methods have been used for other tasks like pose detection [23] [24] [25]. For facial landmark detection, [9] used two smaller facial landmark detection tasks to reduce the effect of rigid transformation, then uses Hourglass [23] backbone to obtain the final results. Look at Boundary (LAB) [10] generates a set of boundary heatmap, and uses them as a hint of regressing facial landmarks. [11] kept the high-resolution representation of the input throughout the entire network, directly predict facial landmarks in an end-to-end framework.

3. Methodology

With inspiration from the global regression network in [14], the method proposed in this paper made a derivation from ResNet-18 [26] by changing the average pooling layer to a depth-wise convolutional layer. Such changes is shown empirically to significantly boost the performance of the model significantly compared to using ResNet-18 [26] network directly.

4. Experiments

4.1 Datasets

300W [12] is a combination of four different facial landmark datasets, HELEN [27], LFPW [28], AFW [29] and IBUG. Each image are labeled with 68 landmark points. Its training set consists of 3148 images, the training set of HELEN and LFPW, and all of AFW. Its 689 test images are split to a common part consists of 444 images from the test set of HELEN and LFPW, and a challenging part consists of 135 images of IBUG. The challenging subset is one of the most difficult facial landmark dataset, with significant pose and expression variations. Its official test set [30], used for competition, consists of 300 indoor and 300 outdoor labeled images taken in the wild.

COFW [13] is a dataset that focusses on occlusion. It has 1345 training images and 507 test images, labeled with 29 facial landmark points in the same format as in LFPW [28]. 845 of the training images are from LFPW [28].

The training and testing was only done on the 300W dataset. In a latter stage, experiments will be done on other common datasets like COFW.

4.2 Experiment Settings

4.2.1 Data Augmentation

Different data augmentation techniques are used. The images and labels are randomly rotated by [−30, 30] degrees before cropping to the bounding box. Also, samples are flipped with a probability of 0.5. By doing this, the labels symmetric to each other, like the ones for left eye and right eye, also switches position. The augmentations are performed with the PIL package, and are done online during training. In each epoch, different random samples are generated.
4.2.2 Training Configuration
The models are trained using PyTorch. Training are done using mini-batches with batch size 50. The Adam Optimizer is used with an initial learning rate of 0.001. Learning rate decays to half of the original if no improvement were obtained for 32 consecutive epochs with this learning rate. The learning rate is allowed to decay a maximum of 10 times. The $l^2$ loss is used throughout training.

4.3 Evaluation Metric
The Normalized Root Mean Squared Error (NRMSE) [31] is used for evaluation. To compare the result of this method with other pre-existing methods fairly, inter-pupil and inter-ocular normalization are used. The pupil location of an eye is calculated by averaging all six landmark points around the eye for 300W dataset.

4.4 Results and Discussion
The results on 300-W dataset are reported in Table 1. Some faces that generated the most errors are shown in Figure 1.

Since the proposed method does not use cascading structure, which has been proven to improve accuracy [2], the performance is lower than some networks utilizing both deep CNNs and cascading structures. Comparing with GRegNet [14], the author claimed a significantly higher performance on the GRegNet compared to the method proposed by this paper, on par with state-of-arts by then, despite some similarities of architectures. That method utilizes a large number of pixel-based data augmentation, including but not limited to blurring, noise and grayscale. However, their augmented data is also generated online similar to the data generation method proposed in this paper. They claimed that their model is trained for up to 3000 epochs with a patience of 200 epochs, while

### Table 1: Comparison of model performance (NRMSE; $\times 10^{-2}$) on 300-W

| Method          | Common | Challenging | Fullset |
|-----------------|--------|-------------|---------|
| Inter-ocular normalization |        |             |         |
| PCD-CNN [32]    | 3.67   | 7.62        | 4.44    |
| SAN[33]         | 3.34   | 6.60        | 3.98    |
| LAB [10]        | 2.98   | 5.19        | 3.49    |
| GRegNet [14]    | 2.97   | 5.07        | 3.38    |
| HRNetV2-W18 [11]| 2.87   | 5.15        | 3.32    |
| LRefNet[14]     | 2.71   | 4.78        | 3.12    |
| Proposed Method | 3.65   | 6.42        | 4.19    |
| Inter-pupil normalization |        |             |         |
| RCPR [13]       | 6.18   | 17.26       | 8.35    |
| CFAN [34]       | 5.50   | 16.78       | 7.69    |
| ESR [7]         | 5.28   | 17.00       | 7.58    |
| SDM [35]        | 5.57   | 15.40       | 7.50    |
| LBF [8]         | 4.95   | 11.98       | 6.32    |
| CFSS [36]       | 4.73   | 9.98        | 5.76    |
| 3DDFA [37]      | 6.15   | 10.59       | 7.01    |
| TCDCN[6]        | 4.80   | 8.60        | 5.54    |
| MDM [38]        | 4.83   | 10.14       | 5.88    |
| RAR [39]        | 4.12   | 8.35        | 4.94    |
| DVLN[40]        | 3.94   | 7.62        | 4.66    |
| GRegNet [14]    | 4.11   | 7.32        | 4.74    |
| LRefNet[14]     | 3.76   | 6.89        | 4.37    |
| LAB [10]        | 3.42   | 6.98        | 4.12    |
| Proposed Method | 5.07   | 9.28        | 5.89    |
Figure 1: Some faces with high test error. Blue points are the landmarks extracted by the proposed work. Green points are the ground truth.

for the proposed method, the loss stopped decreasing after no more than 300 epoches. This might be able to attribute to pose distribution balancing being used in that work. Even though, it is unclear whether their reported performance of GRegNet is from a model also trained with LRefNet concurrently. Future works including implementing a more complex structure based on the current network, and also test it on more
datasets.

5. Conclusion
This paper have attempted to reimplement a part of the recent [14], and made experiments. Detailed comparison with the network the model is trying to reimplement is done. Although the current results are not as good as the original work this paper tried to reimplement, those could be attributed to the original author using suspectable training configurations.

References
[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.
[2] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013.
[3] T. Cootes, C. Taylor, D. Cooper, and J. Graham, “Active shape models-their training and application,” Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38–59, 1995.
[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 681–685, June 2001.
[5] D. Cristinacce and T. Cootes, “Automatic feature localisation with constrained local models,” Pattern Recognition, vol. 41, no. 10, pp. 3054–3067, 2008.
[6] 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, vol. 38, pp. 918–930, May 2016.
[7] X. Cao, Y. Wei, F. Wen, and J. Sun, “Face alignment by explicit shape regression,” International Journal of Computer Vision, vol. 107, pp. 177–190, Apr 2014.
[8] S. Ren, X. Cao, Y. Wei, and J. Sun, “Face alignment at 3000 fps via regressing local binary features,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.
[9] J. Yang, Q. Liu, and K. Zhang, “Stacked hourglass network for robust facial landmark localisation,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017.
[10] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, “Look at boundary: A boundary-aware face alignment algorithm,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
[11] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, “High-resolution representations for labeling pixels and regions,” arXiv preprint arXiv:1904.04514, 2019.
[12] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge,” in The IEEE International Conference on Computer Vision (ICCV) Workshops, June 2013.
[13] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, “Robust face landmark estimation under occlusion,” in The IEEE International Conference on Computer Vision (ICCV), December 2013.
[14] J. Su, Z. Wang, C. Liao, and H. Ling, “Efficient and accurate face alignment by global regression and cascaded local refinement,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.
[15] T. F. Cootes and C. J. Taylor, “Active shape models — ‘smart snakes’,” in BMVC92 (D. Hogg and R. Boyle, eds.), (London), pp. 266–275, Springer London, 1992.
[16] S. Milborrow and F. Nicolls, “Locating facial features with an extended active shape model,” in Computer Vision – ECCV 2008 (D. Forsyth, P. Torr, and A. Zisserman, eds.), (Berlin, Heidelberg), pp. 504–513. Springer Berlin Heidelberg, 2008.

[17] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, and M. Pantic, “Generic active appearance models revisited,” in Computer Vision – ACCV 2012 (K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, eds.), (Berlin, Heidelberg), pp. 650–663, Springer Berlin Heidelberg, 2013.

[18] R. Gross, I. Matthews, and S. Baker, “Generic vs. person specific active appearance models,” Image and Vision Computing, vol. 23, no. 12, pp. 1080 – 1093, 2005.

[19] E. Antonakos, J. Alabort-i Medina, G. Tzimiropoulos, and S. Zafeiriou, “Hog active appearance models,” in 2014 IEEE International Conference on Image Processing (ICIP), pp. 224–228, Oct 2014.

[20] G. Tzimiropoulos and M. Pantic, “Optimization problems for fast aam fitting in-the-wild,” in The IEEE International Conference on Computer Vision (ICCV), December 2013.

[21] J. M. Saragih, S. Lucey, and J. F. Cohn, “Deformable model fitting by regularized landmark mean-shift,” International Journal of Computer Vision, vol. 91, pp. 200–215, Jan 2011.

[22] N. Kumar, P. Belhumeur, and S. Nayar, “Facetracer: A search engine for large collections of images with faces,” in Computer Vision – ECCV 2008 (D. Forsyth, P. Torr, and A. Zisserman, eds.), (Berlin, Heidelberg), pp. 340–353, Springer Berlin Heidelberg, 2008.

[23] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in Computer Vision-14th European Conference, ECCV 2016, Proceedings, pp. 483–499, Springer Verlag, 2016.

[24] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, “Learning feature pyramids for human pose estimation,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 1281–1290, 2017.

[25] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, “Multi-scale structure-aware network for human pose estimation,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 713–728, 2018.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[27] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, “Interactive facial feature localization,” in Computer Vision – ECCV 2012 (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), (Berlin, Heidelberg), pp. 679–692, Springer Berlin Heidelberg, 2012.

[28] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, “Localizing parts of faces using a consensus of exemplars,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 2930–2940, Dec 2013.

[29] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886, June 2012.

[30] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: database and results,” Image and Vision Computing, vol. 47, pp. 3 – 18, 2016. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.

[31] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, “Robust facial landmark detection via occlusion-adaptive deep networks,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[32] A. Kumar and R. Chellappa, “Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[33] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, “Style aggregated network for facial landmark
detection,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] J. Zhang, S. Shan, M. Kan, and X. Chen, “Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment,” in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 1–16, Springer International Publishing, 2014.

[35] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[36] S. Zhu, C. Li, C. Change Loy, and X. Tang, “Face alignment by coarse-to-fine shape searching,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[37] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, “Face alignment across large poses: A 3d solution,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[38] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[39] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, “Robust facial landmark detection via recurrent attentive-refinement networks,” in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 57–72, Springer International Publishing, 2016.

[40] W. Wu and S. Yang, “Leveraging intra and inter-dataset variations for robust face alignment,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.