Facial Landmark Detection Using Cascade and Direct Regression

Aidan Brasseur
Department of Computer Science
University of Toronto
aidan.brasseur@mail.utoronto.ca

Priyanshu Arora
Department of Computer Science
University of Toronto
priyanshu.arora@mail.utoronto.ca

Abstract

Building upon the fundamental three-level cascade convolutional network proposed by Sun et al [6] to perform facial landmark detection, we offer a direct regression convolutional network architecture which merges the three levels of independently trained neural networks into a single network which can be trained as one. Using a train and test dataset composed of human faces compiled by Sun et Al [6], we trained two models, one closely modeling the cascade approach of [6], and the second using our modified direct regression approach, and found that the direct regression approach offers far lower generalization loss and visually noticeable improvement in facial landmark predictions. While our motivation was that this method would leverage the region-specific fine tuning of the cascade regression architecture while also allowing information to be propagated back to improve the initial coarse facial landmark estimation, this turned out not to be the explanation for the performance improvement.

1 Introduction

Facial landmark detection is crucial for facial identification, facial expression estimation, and various other applications [5]. In recent years, CNN's have shown promising results in solving this task because of the high-level contextual information required for this problem. The CNN-approach can be divided into regression and heatmap approaches. In our research we will focus on the regression approach which can be further subdivided into direct regression and cascade regression. For human faces, cascade regression has been successful because of its coarse-to-fine strategy, starting with an initial rough estimation of the facial landmarks and then iteratively refining the prediction. A study at the Chinese University of Hong Kong found success in their Deep Convolutional Neural Network Cascade [6]. Their algorithm aims to detect 5 facial landmark points: Left eye center, right eye center, nose tip, left mouth corner, and right mouth corner. This study has been a fundamental stepping stone for modern facial landmark detection techniques, and we seek to gain a deeper understanding of this fundamental structure through comparison with a direct regression model created by combining the cascaded levels into one large network.

2 Related Works

Significant progress on facial keypoint detection has been made in recent years after the success of convolutional neural networks. Before Heatmap Regression, Cascade regression was the predominant method in 2D face alignment. Cascade regression found success because of its course-to-fine strategy where an initial prediction is refined over levels of cascading. In the study by Sun et al [6], a deep CNN cascade is created with three levels. The first level consists of 3 deep CNN's, F1, EN1, and NM1, whose input regions cover the whole face (F1), eyes and nose (EN1), and nose and mouth (NM1).

Preprint. Under review.

Each model is trained independently and simultaneously predicts the facial points. For each facial point, the predictions of multiple networks are averaged to reduce the variance. The three networks have the same structure but with different sizes since the sizes of their input regions are different. Networks at the second and third level take local patches centered at the predicted positions of the previous level and are only allowed to make small changes to the previous predictions. Predictions at the last two levels are strictly restricted because local appearance is sometimes ambiguous and unreliable. The networks are only looking at small patch sizes so the networks are shallower than the first level networks. Networks at the first level aim to estimate facial landmark positions based on the whole context of the image, providing robust predictions with few large errors. The networks at the following levels are designed to refine the predictions by only looking at local patches around the initial prediction and achieve high accuracy. [6].

In recent years, Heatmap regression has revolutionized pose estimation. Heatmap regression employs end to end training and requires little hand engineering [1]. Bulat et al [1] created Face alignment NetworkFAN, a deep convolutional neural network comprised of 4 stacked Hour Glass networks. Bulat et al argues that 2d face alignment is a solved problem.

Dollár et al [3] employed cascade regression using a series of weak random fern regressors to estimate the pose of 2D objects. Other methods include CNN with heatmaps and PWC (pixelwise classification) [4].

3 Method/Algorithms

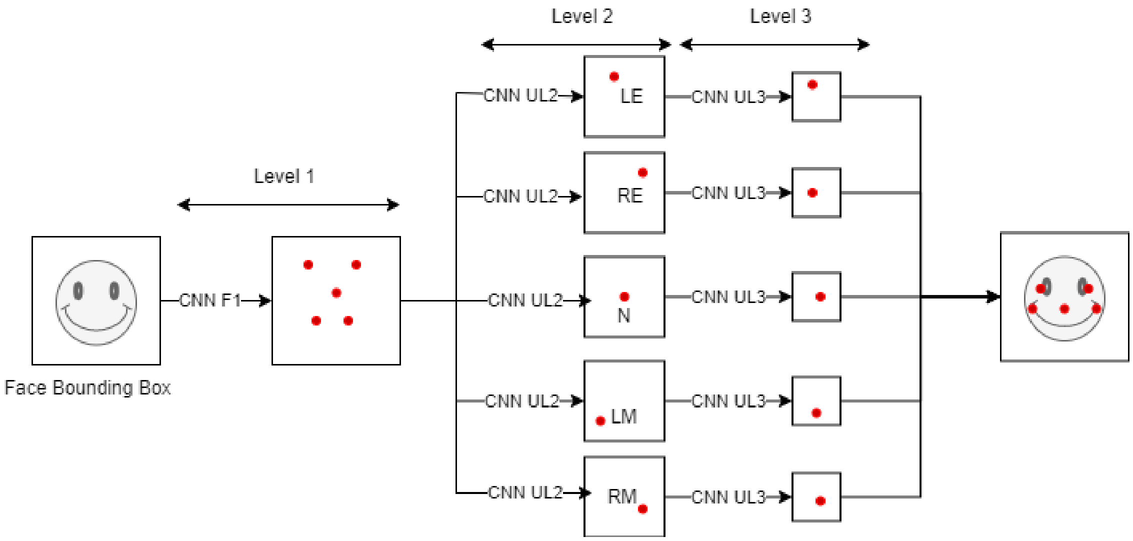

Figure 1: A simplified three-level cascaded convolutional network inspired by Sun et al [6]. The initial input is the facial region returned by a face detector. Level 1 is a single deep convolutional network, denoted F1, which predicts all 5 keypoints. At level two, there are 5 neural networks corresponding to each facial landmark. Each network is fed as input a small 32x32 window around the prediction of F1. Similarly, at level 3 there are 5 neural networks corresponding to each facial landmark, with each receiving a small 16x16 window around the corresponding prediction of level 2.

Table 1: Cascaded Convolutional Network Structure

	Layer 0	**Layer 1**	**Layer 2**	**Layer 3**	**Layer 4**	**Layer 5**	**Layer 6**	**Layer 7**	**Layer 8**	**Layer 9**
F1	I(64, 64)	C(4, 20)	P(2)	C(3, 40)	P(2)	C(3, 60)	P(2)	C(2, 80)	F(120)	F(10)
UL2	I(32, 32)	C(4, 20)	P(2)	C(3, 40)	P(2)	F(60)	F(2)			
UL3	I(16, 16)	C(4, 20)	P(2)	C(3, 40)	P(2)	F(60)	F(2)			

F1 is the structure of the first level neural network, UL2 is the structure of the second level neural networks, and UL3 is the structure of the third level neural networks. C(k, n) denotes a convolutional layer with kernel size k and n feature maps, P(k) denotes a max pooling layer with kernel size k, and

F(n) denotes a fully connected layer with output n.

Algorithm 1: Forward Method

```
Function Forward($input$):
    level1landmarks = F1(input)
    for i in number of landmarks do
        window32 = create 32 x 32 window around level 1 landmark i

        level 2 landmarks = UL2(window32)
        level 2 landmarks = convert landmark predictions into the 64x64 image coordinate space

        window16 = create 16x16 window around level 2 landmark i
        level 3 landmarks = UL3(window16)
        level 3 landmarks = convert landmark predictions into the 64x64 image coordinate space

        final prediction = weighted average between predictions of level1, level2, and level3
    end
    return final predictions
```

Both the cascade regression and direct regression approaches share the same structure shown above, including the same forward method, with the only difference being that the direct regression approach includes the final predicted weighted average weights as learnable parameters. Where the models differ is in the way that we train them.

For the cascade regression model, we trained all of the component neural networks independently. This entailed training the F1 neural network on the input data and training all 5 level 2 and level 3 networks by creating a 32 x 32 or 16x16 window around the training target landmark and then randomly shifting the window before training the network. In comparison, for the direct regression approach we train all the networks together using a single loss function propagated backwards.

The loss function we are using in both cases is the mean square error formula:

$$\mathcal{L}(x, y) = \frac{1}{N} \sum_{n=1}^{N} (x_n - y_n)^2$$

4 Experiments

The dataset we used during our training and testing was sourced by Sun et al [6], and is composed of 13 466 facial images from the LFW dataset and the web. To simplify our training procedure, we decided to use only a small subset of this dataset with 2048 images during training. Our test set is also a very small subset of the dataset at just 256 images to simplify our comparison.

Comparing the performance of the two models, we see that the direct regression approach achieves generalization error of 4.22, while the cascade regression model had a larger generalization error of 9.57. This performance improvement can also be seen visually in Figure 2, with the end-to-end predicted landmarks being noticeably closer to the ground truth targets.

While this performance improvement is as we expected, it did not perform in the way that we expected. Since end-to-end training allows the early cascade level CNNs to be supplemented by information from later cascade level CNNs through backpropagation, we expected that in the end-to-end approach the F1 layer would be able to improve its prediction after receiving fine-tuning from the level 2 and level 3 networks. However, we observed after comparing the performance of the F1 network after end-to-end training compared to the F1 network which was independently trained that the end-to-end network achieved higher generalization loss of 15.93 compared to a loss of 11.87 when independently trained.

The first and third samples from Figure 3 suggest that our model has a bias towards a forward facing facial structure where the eyes and mouth corners are symmetrically spaced away from the nose. These samples deviate from this structure because the face is angled slightly away from the camera, breaking the symmetry that the model is biased towards predicting. This suggests that our training data contains many more faces facing directly towards the camera than faces angled away, and this

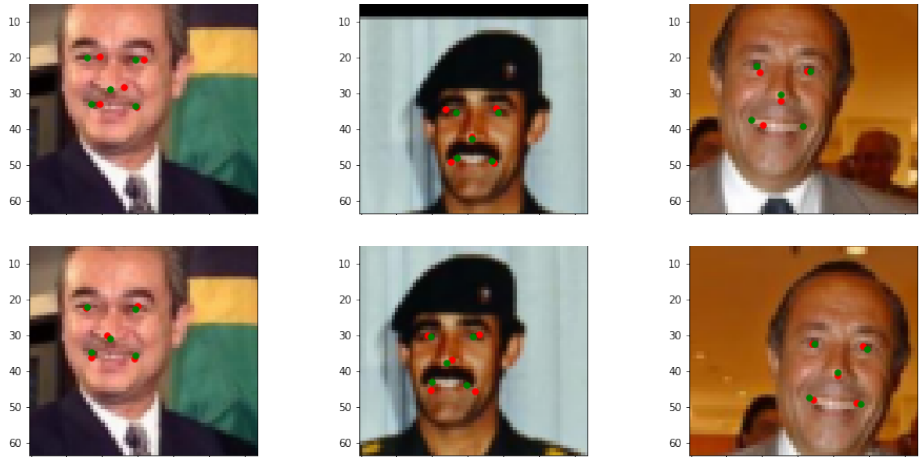

Figure 2: Comparison of Cascade Regression (Top) to End-to-End model (Bottom): Green is ground truth and red is predicted landmarks

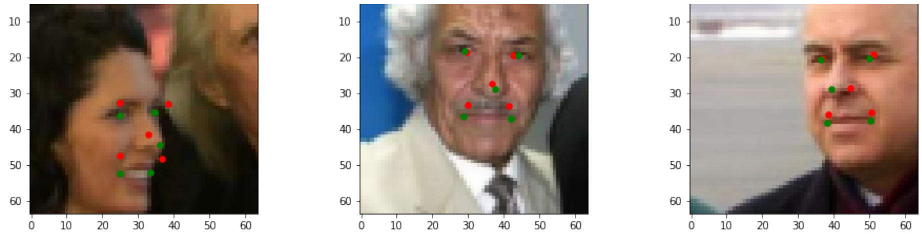

Figure 3: Examples of failure points for the end-to-end model: Green is ground truth and red is predicted landmarks

could perhaps be rectified by using a larger subset of the data. The second sample in Figure 3 shows that our model is susceptible to obstructions in the face, in this case with the model believing that the moustache is the mouth. We see this occur often where obstructions from moustaches, hair, glasses, and wrinkles in the face around the mouth causes the model to make incorrect predictions.

In the cascade, all models are trained independently, so the models at level 2 and level 3 are only looking at a small context and do not have information of the other landmarks. There are two problems that arise from this: The first is that since the context of the window is low, the prediction can be unreliable since small obstructions can be mistaken as the true landmark of interest. The second is that since there is no information of other landmarks, preservation of the overall face structure is lost. For these two reasons, we must weight the prediction of the first level higher than the second level which in turn must be higher than the third level. We can see experimentally that this is verified, as if we were to add more weight to the higher levels, and less to the lower levels, then the predicted points deviate from the overall normal face structure. These weights are hyperparameters that need to be tuned and one of the benefits of the direct regression model is that these weights can be learned so there is no need for tuning. In addition, as mentioned by Cong et al [2], since the backpropagation gives context to all the points, the higher level CNN is supplemented by the information from lower level CNN.

5 Conclusion

We explored an end-to-end training approach to a cascaded regression model and found significant performance improvements when compared to the original independently trained approach of Sun et al [6]. By exploring the benefits of end-to-end training, this serves to justify and understand the shift towards end-to-end techniques as seen in modern facial landmark detection algorithms.

References

[1] Bulat, Adrian & Tzimiropoulos, Georgios. (2017). How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). 10.1109/ICCV.2017.116.

[2] Cong, W., Zhao, S., Tian, H., & Shen, J.. Improved Face Detection and Alignment using Cascade Deep Convolutional Network.

[3] Dollár, P., Welinder, P., & Perona, P. (2010). Cascaded pose regression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 1078-1085).

[4] Hsu, C., Lin, C., Hung, T., Lei, C.L., & Chen, K. (2020). A Detailed Look At CNN-based Approaches In Facial Landmark Detection. ArXiv, abs/2005.08649.

[5] Khabarlak, K., & Koriashkina, L. (2021). Fast Facial Landmark Detection and Applications: A Survey. CoRR, abs/2101.10808. https://arxiv.org/abs/2101.10808

[6] Sun, Y., Wang, X., & Tang, X. (2013). Deep Convolutional Network Cascade for Facial Point Detection. In 2013 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3476-3483).

A Contributions

Since we are roommates, all code was pair programmed together, and the paper was written together as well with equal contribution.