Disentangling 3D-Pose in A Dendritic CNN for Unconstrained 2D-Face Alignment - Supplementary

Anonymous CVPR submission

Paper ID 1230

Dataset	Pre-Aug	Post-Aug
AFLW-PIFA (PCD-CNN-Fast)	2.85	**2.81**
AFW (PCD-CNN-Fast)	2.80	**2.66**
AFLW-PIFA (PCD-CNN-C+C)	2.49	**2.40**
AFW (PCD-CNN-C+C)	2.52	**2.36**
COFW (PCD-CNN-Fast)	6.02	**5.77**
300W-Challenge (PCD-CNN-Fast)	7.62	**7.17**

Table 1: NME on different datasets Pre-Augmentation and Post-Augmentation during testing.

1. Improvement in localization by augmentation during testing

For a fair comparison with the previous state-of-the-art methods we did not perform augmentation during testing. In the next set of experiments along with the test image, we also pass the flipped version of it and the final output is taken as the mean of the two outputs. With experimentation we observe that data augmentation while testing also improves the localization performance. This does not incur any increase in run-time as the inputs can be passed through the network in batch mode, keeping the runtime still at 20FPS. Table 1 shows the effects of data augmentation during testing.

2. Effect of Pose Disentaglement

Next, we also perform an experiment to observe the effect of 3D pose conditioning on the second auxiliary network designed for fine grained localization. Table 2 shows the effect of disentangling pose by conditioning, when the auxiliary conv-deconv network does not receive information from the PoseNet.

3. Magnified version of the Tree

One expects to receive information from all other keypoints in order to optimize the features at a specific keypoint. However, this has two drawbacks: First, to model the interaction between keypoints lying far away such as 'eye corner' and 'chin', convolution kernels with larger size have to be introduced. This leads to increase in the number of parameters. Secondly, relationships between some keypoints are unstable, such as 'left eye corner' and 'right eye corner'. In a profile face image one of the points may not be visible and passing information between those two keypoints may lead to erroneous results. Hence, convolution kernels are learned at the size of 14×14 which ensures keypoints which are closer and have stable relationships to be connected together.

Method	NME
PCD-CNN + Auxiliary Network	2.99
PCD-CNN + Pose Conditioned Auxiliary Network	2.49

Table 2: Mean square error normalized by bounding box calculated on AFLW test set following PIFA protocol. When PCD-CNN and fine-grained localization network both are conditioned on pose yields lower error rate.

We also describe the process of extending the proposed dendritic structure of facial landmarks to other datasets with variable number of landmark points. Figure 4a shows the tree structure of the 21 landmark points compatible with the AFLW dataset. In figure 4b and 4c the number of points is increased to 29 and 68 respectively compatible with COFW and 300W datasets. We wish to keep the structure of the facial landmarks intact while increasing the number of landmark points. For this, we make use of the network surgery. First, the number of deconvolution filters in the penultimate and ultimate deconvolution layers is increased to 128 and 64 respectively. Next 1×1 convolutions are used to obtain desire number of outputs, which is then sliced and concatenated in order for loss computation. For instance, eye center points is split into 4 landmark points in the case of COFW and 300W datasets, and ear corner points are dropped. An advantage of network surgery is that, it leads to yielding a variable number of landmark points with minimal increase in parameters while keeping the face structure intact.

4. Training Details

KeypointNet and PoseNet described in section 3 are designed based on the SqueezeNet architecture, attributing its lower parameter count. The proposed PCD-CNN was first trained using AFLW training set, where Mask-Softmax is used for keypoints and Euclidean Loss for 3D pose estimation. Starting from the learning rate of 0.001, the network was trained for 10 epochs with momentum set to 0.95. The learning rate was dropped by a factor of 10 every 3 epochs. While training PCD-CNN for COFW and 300W datasets, the convolution branch was initialized with the previously trained network, whereas the deconvolution branches were trained from scratch. Since, COFW and 300W datasets does not provide 3D pose ground truth, we leverage the previously trained PoseNet and freeze its weights. As shown in the section 3 of the main paper, disentangling pose by conditioning improves the localization performance.

4.1. Training PCD-CNN for COFW

This section covers the details of training for the COFW dataset. The PCD-CNN network was trained using the Mask Softmax and hard negative mining. The second auxiliary network was trained for the task of occlusion detection. According to the released details about the COFW dataset, around 23% of the landmark points are invisible. Hence, to tackle the class imbalance problem between the visible and invisible points the following loss function was used.

$$L(\boldsymbol{p},\boldsymbol{g}) = \sum_{i=1}^{29}(0.23*\mathbb{1}_{g_i^{vis}=1}+0.77*\mathbb{1}_{g_i^{vis}=0})(p_i^{vis}-g_i^{vis})^2 \quad (1)$$

where $\boldsymbol{p},\boldsymbol{g}$ are the vector of predicted and ground-truth visibilities. p_i^{vis} and g_i^{vis} are the values of the individual elements in the vectors of visibilities. The weighted loss function also balances the gradients back-propagated while loss calculation.

Figure 1 shows the failure rate and error rate on the COFW dataset. The failure rate on the COFW dataset drops to 4.53% bringing down the error rate to 6.02. When testing with the augmented images the error rate further drops to 5.77 bringing it closer to human performance 5.6. Figure 3a shows the precision recall curve for the task of occlusion detection on the COFW dataset. PCD-CNN achieves a significantly higher recall of 44.7% at the precision of 80% as opposed to RCPR's [1] 38.2%.

5. Hard mining

Figure 2 shows the distribution of average normalized error on the training sets of AFLW and COFW datasets. The error distributions were obtained upon evaluating the PCD-CNN network on the training set, after it is trained with the whole dataset for 10 epochs. The dataset is partitioned into hard and easy samples after choosing the mode of the distribution as the threshold. Next, the network is trained again, by sampling equal number of images from both groups, which results in an effective reuse of the hard examples.

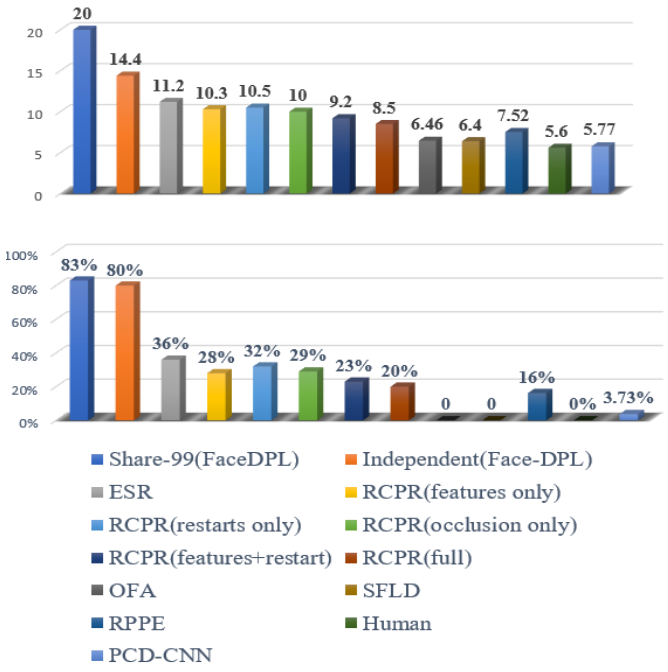

Figure 1: Comparison of NME and failure rate over visible landmarks out of 29 landmarks from the COFW dataset.

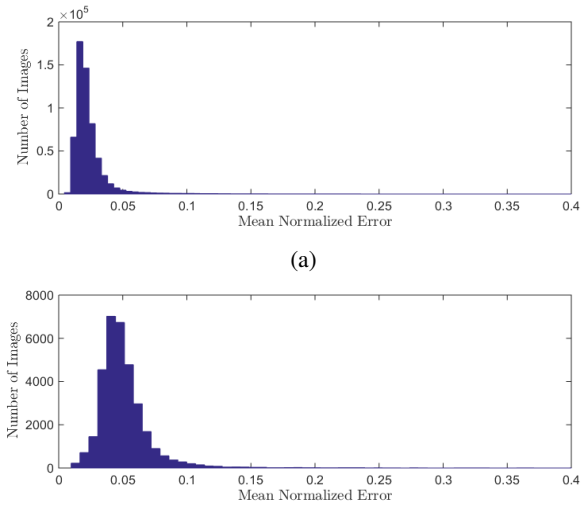

Figure 2: Histogram of error, when evaluated on the training set of (a) AFLW (b) COFW.

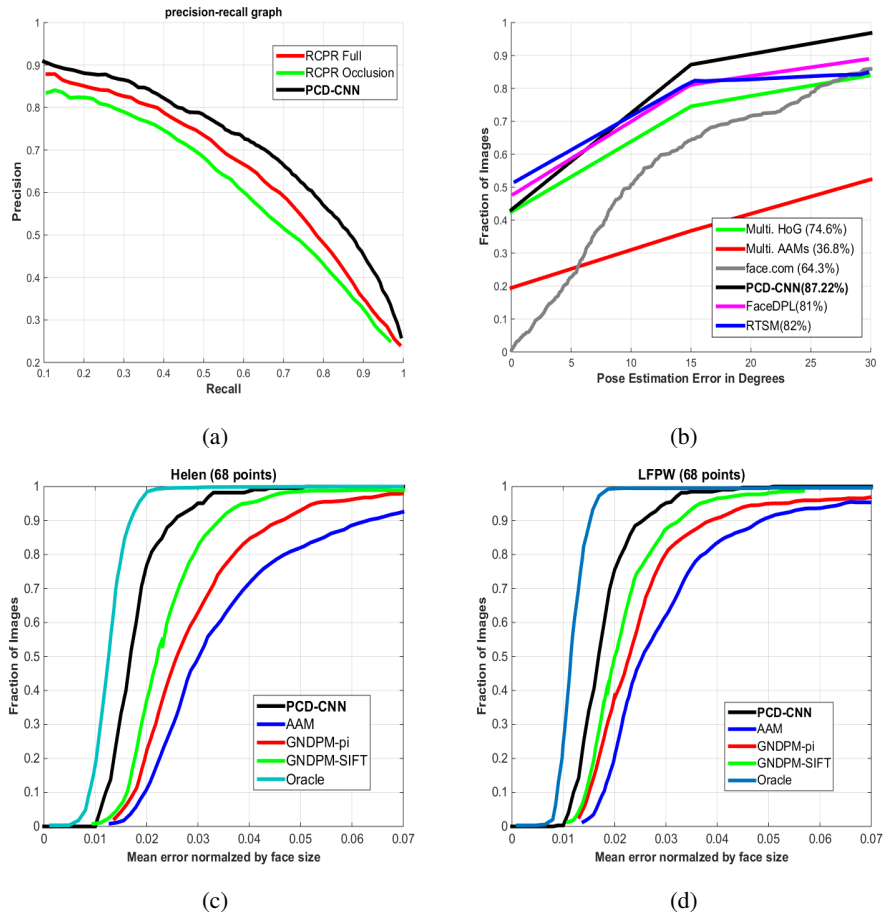

Figure 3: (a) Precision Recall for the occlusion detection on the COFW dataset. (b)Cumulative error distribution curves for pose estimation on AFW dataset. The numbers in the legend are the percentage of faces that are labeled within $\pm 15^{\circ}$ error tolerance. Cumulative Error Distribution curve for (c) Helen (d) LFPW, when the average error is normalized by the bounding box size.

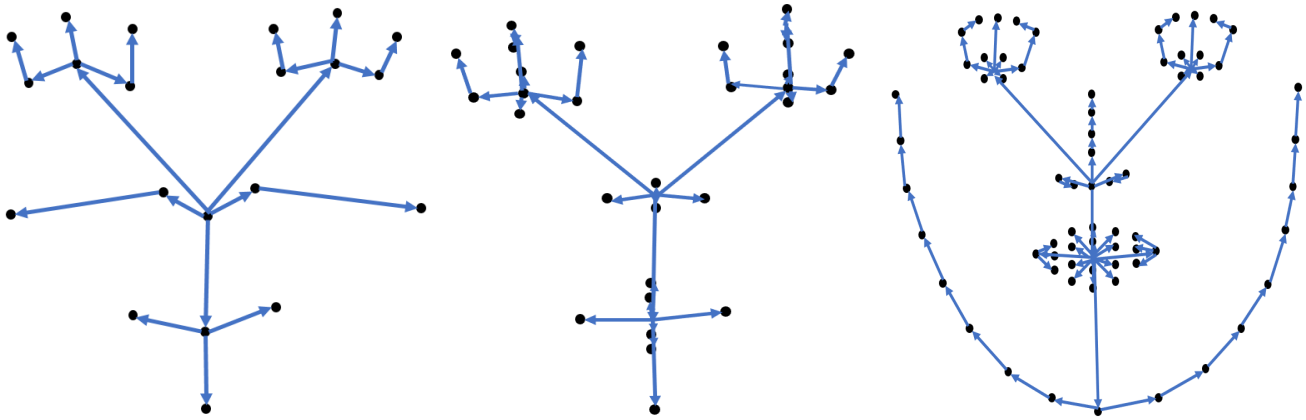

Figure 4: The proposed extension of the dendritic structure from Figure 1 of the main paper, generalizing to other datasets with variable number of points.

Figure 5: Qualitative results generated from the proposed method. The green dots represent the predicted points. Every two show randomly selected samples from AFLW, AFW, COFW, and 300W respectively with all the visible predicted points.

6. More results on AFLW, AFW, LFPW and HELEN

In this section, we show some more results obtained by the PCD-CNN on AFW, LFPW and Helen datasets. Figure 3b shows the cumulative error distribution curves for the prediction of face pose on AFW dataset. We observe that even though the primary objective of PCD-CNN is not pose prediction, it achieves state-of-the-art results when compared to recently published works Face-DPL [4],RTSM [2].

Figures 3c and 3d show the cumulative error distribution curve on LFPW and Helen datasets, when the average error is normalized by face size. PCD-CNN achieves significant improvement over the recent work of GNDPM [3].

Figure 5 shows some of the difficult test samples from AFLW, AFW, COFW and IBUG datasets respectively.

References

[1] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. *ICCV*, 0:1513–1520, 2013. 2

[2] G.-S. Hsu, K.-H. Chang, and S.-C. Huang. Regressive tree structured model for facial landmark localization. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 5

[3] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *CVPR*, pages 1851–1858, June 2014. 5

[4] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, June 2012. 5