Facial Landmarks Detection using a Cascade of Recombinator Networks

Pedro Diego López, Roberto Valle, and Luis Baumela

Univ. Politécnica Madrid, Spain.
diego.lopez.maroto@alumnos.upm.es
{rvalle,lbaumela}@fi.upm.es

Abstract. Nowadays, Convolutional Neural Nets (CNNs) have become the reference technology for many computer vision problems, including facial landmarks detection. Although CNNs are very robust, they still lack accuracy because they cannot enforce the estimated landmarks to represent a valid face shape.

In this paper we investigate the use of a cascade of CNN regressors to make the set of estimated landmarks lie closer to a valid face shape. To this end, we introduce CRN, a facial landmarks detection algorithm based on a Cascade of Recombinator Networks. The proposed approach not only improves the baseline model, but also achieves state-of-the-art results in 300W, COFW and AFLW that are widely considered the most challenging public data sets.

Keywords: Face alignment, Cascaded Shape Regression, Convolutional Neural Networks

1 Introduction

Facial landmarks detection is a fundamental problem in computer vision with applications in many real-world tasks such as attributes and pose estimation [1], facial verification [8], etc. Current state-of-the-art methods are based on deep Convolutional Neural Nets (CNNs). Lv *et al.*'s [7] approach uses CNNs to set up a global and a set of local face parts regressors for fine-grained facial deformation estimation. Xiao *et al.* [10] is one of the first approaches that fuse the feature extraction and regression steps into a recurrent neural network trained end-to-end. Kowalski *et al.* [5] and Yang *et al.* [11] are among the top performers in the Menpo competition [12]. Both use global similarity transform to normalize landmark locations followed by a VGG-based and a Stacked Hourglass network respectively to regress the final shape.

CNN approaches are very robust to face deformations and pose changes due to the large receptive fields of deep nets. However, they lack accuracy because of two factors. First, the loss of feature maps resolution in the concatenation of many convolutional and pooling layers. Second, the difficulty in imposing a valid face shape on the set of estimated landmark positions.

The Recombinator Network addresses the first factor by combining features computed at different scales [3]. This is achieved by processing the image in a set of branches at different resolutions. Finer and deeper branches pass information to the coarser ones allowing for the net to combine the information at different levels of abstraction and scales.

In this paper we address the issue of making the set of estimated landmarks look like a valid face. To this end we present a method called Cascade of Recombinator Networks (CRN) that uses cascade of deep models to enforce valid face shapes on the set of estimated landmark positions. We also introduce a new loss function robust to missing landmarks and an aggressive data augmentation approach to improve Honari *et al.*'s [3] baseline system.

2 Cascade of Recombinator Networks

In this section we present the Cascade of Recombinator Networks (CRN) (see Fig. 1). It is composed of S stages where each stage represents a network that combines features across multiple branches B based on Honari *et al.*'s [3] architecture. The output of each stage is a probability map per each landmark providing information about the position of the L landmarks in the input image. The maximum of each probability map determines the landmarks positions.

The key idea behind our proposal is to employ a cascade of regressors that incrementally refine the location of the set of landmarks. The input for each regressor is the set of probability maps produced by the previous stage of the cascade. Between each cascade stage, we introduce a *map dropout* layer that deletes, with probability p, the map of a landmark (see red-crossed map in Fig. 1). In this way we force the net to learn the structure of the face, since it must predict the position of some landmarks using the location of its neighbors. This idea of ensemble of regressors has been extensively used within the so-called Cascade Shape Regressor (CSR) framework [4, 5, 11].

In our implementation we use a loss function that is able to handle missing landmarks. In this way we can use data augmentation with large face rotations, translations and scalings, labeling landmarks falling outside of the bounding box as missing. It also enables us to train with data sets where some landmarks are not annotated, such as AFLW.

Our loss function, $\mathcal{L}$, is given by

$$\mathcal{L} = \sum_{i=1}^{N} \left(-\frac{1}{||\mathbf{w}_i^g||_1} \sum_{l=1}^{L} \left(\mathbf{w}_i^g(l) \cdot \mathbf{m}_i^g(l) \cdot \log(\mathbf{m}_i(l)) \right) \right) , \qquad (1)$$

where $\mathbf{m}_i(l)$ and $\mathbf{m}_i^g(l)$ represent the predicted probability map and the ground truth respectively, $\mathbf{w}_i^g(l)$ the labeled mask indicator variable (takes value "1" when a landmark is annotated, "0" otherwise), N the number of training images and L the number of landmarks.

We have further improved the accuracy of the Recombinator Network baseline by replacing max-pooling layers with convolutional layers with stride 2.

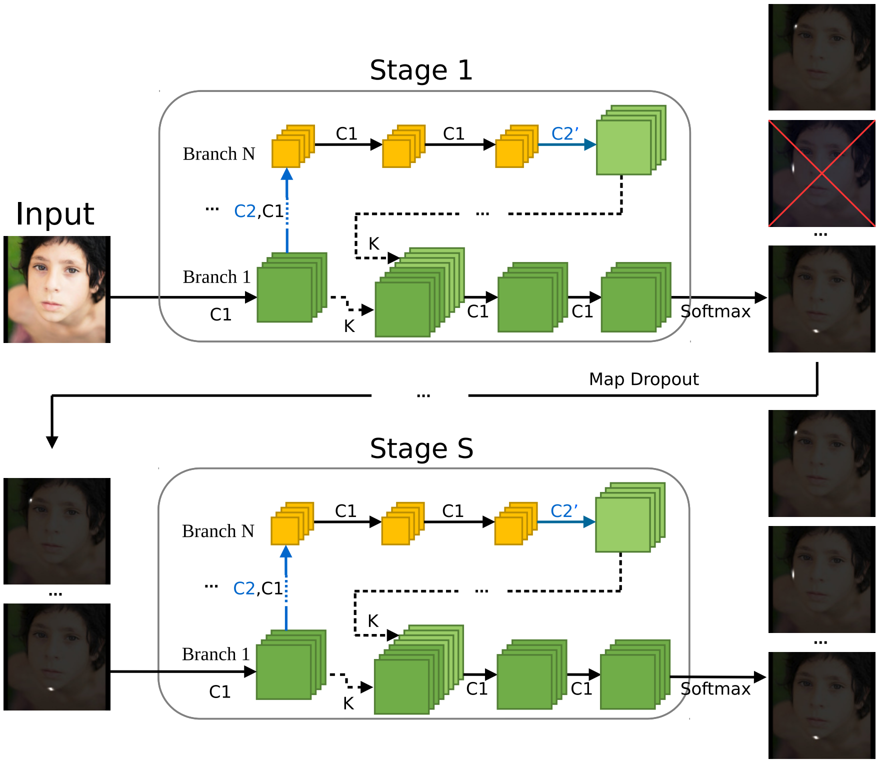


Fig. 1: CRN framework architecture diagram. Each stage is a RCN [3] where $C1$, $C2$ and $C2'$ represent a stride 1 conv layer, stride 2 conv layer and a transpose convolution with stride 2 respectively. The output of each stage is the input to the next one. Between each stage we introduce a *map dropout* layer.

Finally, we found that locating each landmark at the position with maximum probability is very sensitive to noise. We propose to apply a Gaussian smoothing filter to each probability map to improve the robustness of the predictions. Thus, large areas are favored with respect to single pixels with high probability.

3 Experiments

We perform experiments using 300W, COFW and AFLW that are considered the most challenging public face alignment data sets. To train our algorithm we shuffle each training subset and split it into 90% train-set and 10% validation-set.

We use common evaluation metrics to measure the shape estimation error. We employ the normalized mean error (NME), the average euclidean distance between the ground-truth and estimated landmark positions normalized with the constant d_i. Depending on the database we report our results using different values of d_i: the distance between the eye centers (*pupils*), the distance between the outer eye corners (*corners*) and the bounding box size (*height*). The NME is given by

$$NME = \frac{100}{N} \sum_{i=1}^{N} \left(\frac{1}{||\mathbf{w}_i^g||_1} \sum_{l=1}^{L} \left(\frac{\mathbf{w}_i^g(l) \cdot \|\mathbf{x}_i(l) - \mathbf{x}_i^g(l)\|}{d_i} \right) \right), \qquad (2)$$

where $\mathbf{x}_i(l)$ and $\mathbf{x}_i^g(l)$ denote respectively the predicted and ground truth landmarks positions.

In addition, we also use a second group of metrics based on the Cumulative Error Distribution (CED) curve. We calculate AUC_ε as the area under the CED curve for faces with NME smaller than ε and FR_ε as the failure rate representing the percentage of testing faces with error greater than ε.

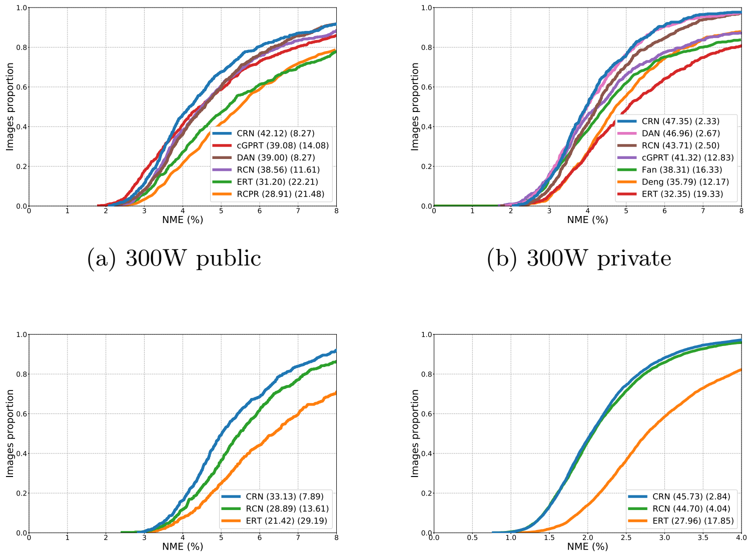
For our experiments we train the CRN stage by stage, selecting the model parameters with lower validation error. We crop faces using the bounding boxes annotations enlarged by 30%. We augment the data in each epoch by applying random rotations between ±30°, scaling by ±15% and translating by ±5% of bounding box size, randomly flipping images horizontally and generating random rectangular occlusions. We use Adam stochastic optimization with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e^{-8}$. We train each stage until convergence. Initial learning rate is $\alpha = 0.001$. When the validation error levels out for 10 epochs, we multiply the learning rate by 0.05. The cropped input face is reduced from 160×160 to 1×1 pixels gradually halving their size across $B = 8$ branches applying a stride 2 convolution with kernel size 2×2[1]. All layers contain 68 filters to describe the required landmarks features. We apply a Gaussian filter with $\sigma = 31$ to the output probability maps to reduce the noise effect. Finally, we set the number of stages $S = 2$ since more stages report a poor improvement. Training using AFLW takes 24 hours using a NVidia GeForce GTX 1080Ti GPU (11GB) with a batch size of 32 images.

At run-time our method requires on average 40 ms to process a detected face, a rate of 25 FPS. This processing speed could be halved reducing the number of CNN stages, at the expense of a slight reduction in accuracy (see CRN (S=1) at Tables 1, 2, 3 and 4).

We compare our model with the top algorithms in the literature. We show in Tables 1, 2, 3 and 4 the results reported in their papers. We have also trained DAN [5], RCN [3], and GPRT [6] with the same settings, including same training, validation and bounding boxes. In Fig. 2 we plot the CED curves. In the legend we provide the AUC_8 and FR_8 values for each algorithm.

From the results in Tables 1 and 2 we can conclude that in the 300W data set our approach provides results with an accuracy comparable to the best in the literature. However, we notice that Yang *et al.* [11] takes several seconds to process one image, whereas ours runs in real-time. In COFW we report the best result in the literature (see Table 3). Similarly, in the largest and most challenging data set, AFLW, we claim to report the best result, since TSR [7]

[1] 5×5 images are reduced to 2×2 pixels applying a kernel size of 3×3.



(a) 300W public

(b) 300W private

(c) COFW

(d) AFLW

Fig. 2: Cumulative error distributions sorted by AUC for each data set.

Method	Common		Challenging		Full			
	pupils	corners	pupils	corners	pupils	corners		
	NME	NME	NME	NME	NME	NME	AUC_8	FR_8
RCN [3]	4.70	-	9.00	-	5.54	-	-	-
RCN+DKM [3]	4.67	-	8.44	-	5.41	-	-	-
DAN [5]	4.42	3.19	7.57	5.24	5.03	3.59	55.33	1.16
TSR [7]	4.36	-	7.56	-	4.99	-	-	-
RAR [10]	4.12	-	8.35	-	4.94	-	-	-
SHN [11]	4.12	-	7.00	4.90	-	-	-	-
CRN (S=1)	4.26	3.07	8.69	6.01	5.09	3.62	55.62	2.75
CRN (S=2)	4.12	2.97	7.90	5.47	4.83	3.44	57.44	1.88

Table 1: Error of face alignment methods on the 300W public test set.

Method	Indoor corners			Outdoor corners			Full corners		
	NME	AUC_8	FR_8	NME	AUC_8	FR_8	NME	AUC_8	FR_8
DAN [5]	-	-	-	-	-	-	4.30	47.00	2.67
SHN [11]	4.10	-	-	4.00	-	-	4.05	-	-
CRN (S=1)	4.42	45.91	1.66	4.45	45.25	2.66	4.43	45.59	2.16
CRN (S=2)	4.28	47.36	2.66	4.25	47.32	2.00	4.26	47.35	2.33

Table 2: Error of face alignment methods on the 300W private test set.

Method	pupils		
	NME	AUC_8	FR_8
RAR [10]	6.03	-	-
Wu *et al.* [9]	5.93	-	-
SHN [11]	5.6	-	-
CRN (S=1)	5.75	30.91	11.04
CRN (S=2)	5.49	33.13	7.88

Table 3: COFW results.

Method	height
	NME
Bulat *et al.* [2]	2.85
CCL [13]	2.72
TSR [7]	2.17
CRN (S=1)	2.29
CRN (S=2)	2.21

Table 4: AFLW results.

ignores the two landmarks attached to the ears, which are the ones with largest error (see Table 4).

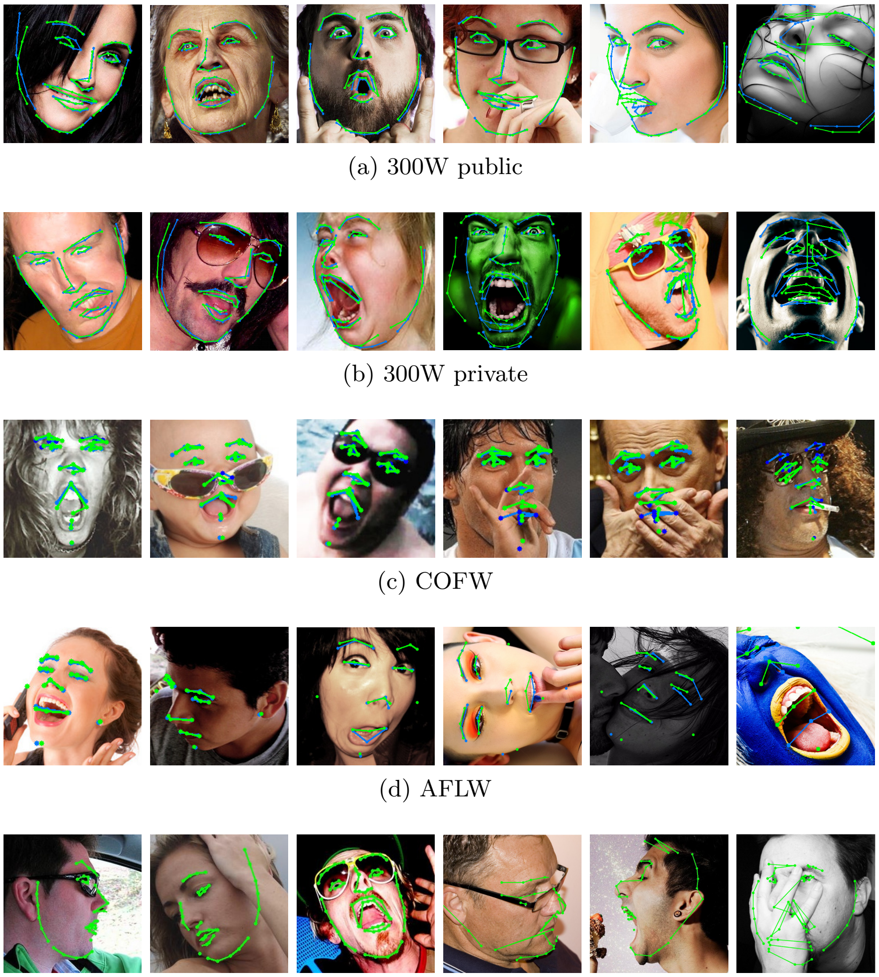
If we consider the CED metrics in Fig. 2, we can see that our approach, CRN, is the one with highest AUC values and smallest FR. In all experiments our CED curve is consistently above the rest except for the cGPRT [6] algorithm in the 300W public data set. In this case, cGPRT reports better results in "easy" faces, with NME below 3.5, and we are much better in the difficult cases, with higher NMEs, and in the final FR_8 and global AUC_8.

We have also compared CRN with the original RCN baseline model and its denoising key-point model approach (RCN+DKM) [3]. Our modifications to the basic net together with the cascade have boosted the result to the top of the state-of-the-art.

Finally, in Fig. 3, we report qualitative results for all data sets. Here we have also included the recent Menpo competition [12] images whose test annotations have not been released.

4 Conclusions

In this paper we have introduced CRN, a facial landmarks detection algorithm that exploits the benefits of a cascade of CNN regressors to make the set of estimated landmark positions lie closer to the valid shape of a human face.



(a) 300W public

(b) 300W private

(c) COFW

(d) AFLW

(e) Menpo

Fig. 3: Representative results using CRN in 300W, COFW, AFLW and Menpo testing subsets. The first three faces and the following three ones show respectively successful and failure cases. Blue and green colors represent ground truth and shape predictions.

We have proved experimentally that our improvements to the basic Recombinator model together with the cascade approach and the data augmentation boost the performance to achieve state-of-the-art results in the 300W data set and the best reported results in COFW and AFLW.

The analysis of the CED curves show that our approach is consistently above all its competitors except for the easy/frontal images in the 300W public set, for which cGPRT [6] has better results. This proves that CNN approaches are more robust in challenging situations, but a standard cascade of regressors with handcrafted local features such as cGPRT may achieve better results when it is properly initialized. To facilitate the reproduction of our results we will release our implementation after publication.

Acknowledgments: Authors acknowledge funding from the Spanish Ministry of Economy and Competitiveness under project TIN2016-75982-C2-2-R.

References

1. Amador, E., Valle, R., Buenaposada, J.M., Baumela, L.: Benchmarking head pose estimation in-the-wild. In: Proc. Iberoamerican Congress on Pattern Recognition (CIARP) (2017)
2. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: Proc. International Conference on Computer Vision (ICCV) (2017)
3. Honari, S., Yosinski, J., Vincent, P., Pal, C.J.: Recombinator networks: Learning coarse-to-fine feature aggregation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
5. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: A convolutional neural network for robust face alignment. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
6. Lee, D., Park, H., Yoo, C.D.: Face alignment using cascade gaussian process regression trees. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
7. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
8. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI) 38, 1997–2009 (2016)
9. Wu, Y., Ji, Q.: Robust facial landmark detection under significant head poses and occlusion. In: Proc. International Conference on Computer Vision (ICCV) (2015)
10. Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.A.: Robust facial landmark detection via recurrent attentive-refinement networks. In: Proc. European Conference on Computer Vision (ECCV) (2016)
11. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
12. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: A step towards the solution. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
13. Zhu, S., Li, C., Change, C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)