Pattern Recognition Letters
journal homepage: www.elsevier.com

Cascade of Encoder-Decoder CNNs with Learned Coordinates Regressor for Robust Facial Landmarks Detection

Roberto Valle[a], José M. Buenaposada[b,**], Luis Baumela[a]

[a]*Departamento de Inteligencia Artificial in Universidad Politécnica de Madrid, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Spain*
[b]*Departamento de Ciencias de la Computación, Arquitectura de la Computación, Lenguajes y Sistemas Informáticos y Estadística e Investigación Operativa in Universidad Rey Juan Carlos, Calle Tulipán s/n, 28933 Móstoles, Spain*

ABSTRACT

Convolutional Neural Nets (CNNs) have become the reference technology for many computer vision problems. Although CNNs for facial landmark detection are very robust, they still lack accuracy when processing images acquired in unrestricted conditions. In this paper we investigate the use of a cascade of Neural Net regressors to increase the accuracy of the estimated facial landmarks. To this end we append two encoder-decoder CNNs with the same architecture. The first net produces a set of heatmaps with a rough estimation of landmark locations. The second, trained with synthetically generated occlusions, refines the location of ambiguous and occluded landmarks. Finally, a densely connected layer with shared weights among all heatmaps, accurately regresses the landmark coordinates. The proposed approach achieves state-of-the-art results in 300W, COFW and WFLW that are widely considered the most challenging public data sets.

1. Introduction

The problem of facial landmark detection aims to estimate the projection of a set of face key points, such as the eye corners, nostrils or the ends of the eyebrows, onto the image plane. It is a fundamental low-level problem in computer vision that enables the extraction of pose invariant information from the image of a face. Thus, potentially improves the performance of relevant tasks such as face recognition (Bhattacharya et al., 2019), facial expressions recognition (Sun et al., 2019) or facial attributes estimation (Bekios-Calfa et al., 2014). Current state-of-the-art methods are based on the consecutive application of different regressors, each trained to refine the prediction of their predecessor. This is the so-called cascaded regression framework (Dollar et al., 2010).

Top performers in recent landmark estimation benchmarks are based on cascading deep Convolutional Neural Nets (CNNs) (Guo et al., 2018; Feng et al., 2018; Yang et al., 2017; Kowalski et al., 2017; Tang et al., 2018). The success of this approach is based on the robustness of deep CNNs to face deformations and extreme pose changes. This is due to the large receptive fields of deep nets. However, there are two factors that decrease their accuracy when processing images taken in unrestricted settings. First, the loss of spatial information as feature maps reduce their resolution in the concatenation of many convolutional and pooling layers. Second, the difficulty in imposing a valid face shape on the set of estimated landmarks. Encoder-decoder nets addresses the first issue by combining features computed at different scales (Honari et al., 2016; Guo et al., 2018; Newell et al., 2016). The second issue is still under investigation. Recent hybrid proposals, based on the combination of a CNN and an Ensemble of Regression Trees (ERT), can learn a prior on 2D face shapes and constitute a promising research direction to address this problem (Valle et al., 2018; Zhang et al., 2018).

In this paper we present a regressor cascade composed of two encoder-decoder CNNs, that robustly estimate the probability distribution of landmark locations (heatmaps), followed by a regressor that estimates the most likely coordinates from the information in the heatmaps. We call this method Cascaded Heatmaps Regression into 2D Coordinates (CHR2C) (see Fig. 1). To increase the regressor accuracy, we train the first CNN to produce a rough estimation of landmark locations.

**Corresponding author.
e-mail: josemiguel.buenaposada@urjc.es (José M. Buenaposada)

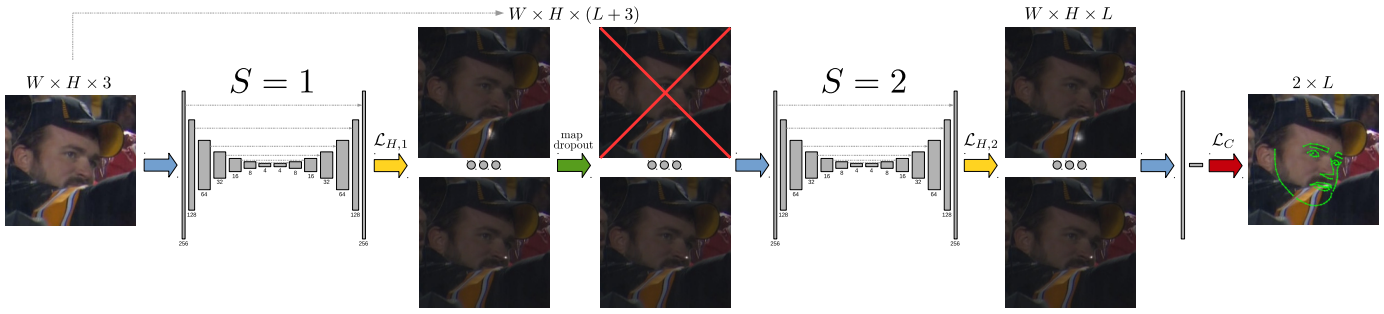


Fig. 1: CHR2C framework architecture diagram. Each stage is an encoder-decoder heatmap regressor with $B = 7$ branches. We show the feature map's width (same as height) in pixels under each network level. Yellow arrows represent the softmax required to produce the heatmaps for the landmarks. Green arrow introduces a *map dropout* layer between each stage (red-crossed map denotes a discarded channel). Finally red arrow represents the regression from heatmaps to 2D coordinates.

The second CNN in the cascade is trained to predict the locations of all landmarks in presence of occlusions and missing feature maps. A preliminary version of our work appeared in López et al. (2018). Here we refine and extend it in several ways. First we introduce a supervision layer between regressor stages to improve the convergence and increase the accuracy of the estimation. We also add a final dense layer with shared weights to regress landmark coordinates from heatmaps and, consequently, an L2 loss function at the output of the last layer. Finally, we also extend the evaluation including the newly released WFLW data base. In the experiments we show that this model achieves top performance results in the most recent benchmarks, 300W, COFW and WFLW. These improvements are most prominent in data bases with a large proportion of occlusions, such as COFW, and extreme poses, expressions and illumination, such as WFLW.

2. Related Work

Facial landmark detection has been a topic of intense research in the computer vision literature for more than twenty years (Wu and Ji, 2019). However, it is still a challenging problem for face images captured "in the wild", in unrestricted settings. Top positions in the most challenging data sets (Sagonas et al., 2016; Burgos-Artizzu et al., 2013; Wu et al., 2018) are taken by cascaded regression methods (Guo et al., 2018; Feng et al., 2018; Yang et al., 2017; Kowalski et al., 2017; Tang et al., 2018). They use a sequence of CNNs to learn an incremental mapping from the raw input image to the final set of face landmark coordinates.

Depending on how the information about the configuration of landmarks in the face is represented, the cascade of CNNs approaches can be organized into two groups: *coordinate regressors* (Xiao et al., 2016; Lv et al., 2017; Kowalski et al., 2017; Yang et al., 2017; Feng et al., 2018), that use a list of 2D landmark coordinates as representation; and *heatmap regressors* (Honari et al., 2016, 2018; Guo et al., 2018; Wu et al., 2018; Tang et al., 2018; Dong et al., 2018), that produce a heatmap per landmark, representing the probability of locating each facial key point at one position in the input image.

The output of each *coordinate regressor* is typically used to rectify the input image before it is further passed to the next step in the cascade. Xiao et al. (2016) fuse the feature extraction and regression steps into a recurrent neural network trained end-to-end. Lv et al. (2017) present a deep regression architecture with two-stage re-initialization to explicitly deal with the initialization problem. Kowalski et al. (2017) and Yang et al. (2017) use a global similarity transform to normalize landmark locations followed by a VGG-based and a Stacked Hourglass network respectively to regress the final shape. Most approaches in this group use L2 loss. Feng et al. (2018) introduce the *wing loss* to pay more attention in the minimization to samples with small errors.

Heatmap regressors use an encoder-decoder fully-convolutional architecture to generate each landmark's heatmap. Honari et al. (2018) designed a network with an equivariant landmark transformation loss to support semi-supervision. Guo et al. (2018) propose to stack dense U-Nets with a novel scale aggregation network topology to achieve accurate results in difficult faces. Wu et al. (2018) complement the Hourglass cascade with a boundary heatmap estimator that provides some shape information and message passing layers to handle occlusions. Tang et al. (2018) is able to achieve good results with quantized densely connected U-Nets with fewer parameters than the stacked Hourglass models (Newell et al., 2016; Yang et al., 2017). Dong et al. (2018) generate images with different styles to increase robustness using a generative adversarial module.

We adopt a heatmap regressor approach in our model, since, compared to a plain list of coordinates, heatmaps provide superior information concerning the landmark's location uncertainty. The architecture of our CNN is similar to RCN (Honari et al., 2016), with some modifications detailed in Sec. 3. Finally, we use a densely connected layer with shared weights among all heatmaps to regress the landmark coordinates. It is different from the fully connected layer of coordinate regressors, *e.g.*, Kowalski et al. (2017); Feng et al. (2018) or the typical `argmax` (Newell et al., 2016; Yang et al., 2017) and `soft argmax` (Honari et al., 2018) in heatmap regressors. In the experiments we show that it contributes to improve the final accuracy in challenging situations.

3. The Proposed Method

In this section we introduce our CHR2C approach (see Fig. 1). It is composed of S stages each made of an encoder-

decoder network that combines features across B branches at different resolutions. Finer and deeper branches pass information to coarser ones allowing the net to combine the information at different levels of abstraction and scales. The output of each stage is a heatmap for each of the L landmarks, providing the probability of each pixel being the actual landmark location in the input image. Moreover, we have developed a simple and effective way of learning how to estimate the 2D coordinates of the corresponding landmarks from the heatmaps.

3.1. Cascaded Heatmaps Regression (CHR)

The key idea behind our proposal is to employ a cascade of regressors that incrementally refines the location of the set of landmarks. The input for each encoder-decoder network is the original input image and the set of heatmaps produced by the previous stage of the cascade. The first stage focuses on learning rigid geometric transformations to roughly estimate the location of visible landmarks. The following stage concentrates on learning the position of occluded landmarks using information about the location of their visible neighbours. Between the two encoder-decoders we introduce a *map dropout* layer that, sets to zero a fraction f of the heatmaps, and a *softmax loss*. The red-crossed heatmap in Fig. 1 means that it has been selected to be "removed" by setting all its values to zero. In this way, the second regressor must learn the relative location of landmarks, since f of them must be predicted from the position of its neighbours.

As shown in Fig. 1 our approach involves two types of loss functions that evaluate the goodness of fit for heatmaps between cascade components and coordinates at the output layer. In addition, these losses are able to handle missing landmarks. This enables us to augment our data with large rigid transformations, treating landmarks falling outside of the bounding box as missing. It also allows us to train the model with data sets having missing landmarks.

We use one-hot encoding for representing the ground truth of each heatmap. Thus, in the ground truth heatmap, $\mathbf{m}_i^g(l)$, we set to 1 the pixel with the l-th landmark location. We employ a softmax to get a sum to one output in the l-th heatmap, $\sum_i^P \mathbf{m}_i(l) = 1$ and adopt the cross-entropy loss for learning the heatmaps,

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

where $\mathbf{w}_i^g$ is a vector with the labeled mask indicator variables for all landmarks ($\mathbf{w}_i^g(l) = 1$ when a landmark is annotated and $\mathbf{w}_i^g(l) = 0$ otherwise), N the number of training images, L the number of landmarks and P the number of pixels.

Similar to Newell et al. (2016), we introduce a heatmap loss head, $\mathcal{L}_H$, between each encoder-decoder module to improve the learning convergence and encourage the output feature maps produced by the decoder to be actual heatmaps.

We set $S = 2$ since more stages produce a small improvement in accuracy and a marked increase in computational cost. We train each module of our system sequentially, followed by an end-to-end refinement. We start with the first heatmap regressor ($S = 1$), trained with extensive rigid data augmentation. Then, using the learned weights as initialization, we cascade the second heatmap regressor ($S = 2$) and train end-to-end including synthetic occlusions (see Fig. 3) and `spatial dropout`, with $f = 0.5$.

To improve the spatial accuracy in our encoder-decoder modules, we only use convolutional layers, replacing RCN max-pooling and up-sampling layers with convolutional and transposed convolutional layers with stride 2. The cropped input face is reduced from 256×256 to 4×4 pixels by gradually halving the spatial extent of feature maps across $B = 7$ branches with stride 2 convolutions. Whenever the spatial resolution is halved we double progressively the number of feature maps, from 64 up to a maximum of 256. We also use batch normalization before the ReLU activation after each convolutional layer.

3.2. Heatmap Regression into 2D Coordinates (HR2C)

Here we introduce an approach for estimating the landmarks coordinates from the heatmaps produced by the CHR net. We term our method HR2C (Heatmap Regression into 2D Coordinates). This is simply a densely connected layer, with shared weights among all heatmaps, regressing the $L \times 2$ landmark coordinates.

Finally, we use an L2 loss to train this layer,

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

where $\mathbf{x}_i(l)$ and $\mathbf{x}_i^g(l)$ represent the l-th landmark predicted and the ground truth coordinates respectively for the i-th training image.

Thus, the loss of the whole model is given by

$$\mathcal{L} = \mathcal{L}_{H,1} + \mathcal{L}_{H,2} + \alpha\cdot\mathcal{L}_C, \quad (3)$$

where $\mathcal{L}_{H,s}$ denotes the heatmap loss $\mathcal{L}_H$ at the output of the s-th encoder-decoder, and α is a weighting parameter balancing the contribution of L2 and softmax losses.

Once we add the HR2C module, we initialize CHR2C with the weights of the CHR and train it end-to-end minimizing (3).

4. Experiments

In this section, we evaluate our algorithm (CHR2C) and the CHR module with one ($S = 1$) and two ($S = 2$) stages in the cascade (see representative results in Fig. 2). We first compare our approach to other methods with public implementation available. In the second set of tests we just compare our results with those reported in the literature for each data set.

4.1. Data sets

In our tests with use 300W, COFW and WFLW, the most challenging public data sets:

- **300W** provides 68 manually annotated landmarks (Sagonas et al., 2016). We follow the established approach and divide the 300W annotations into 3148 training and 689 testing images (public competition). Evaluation is also

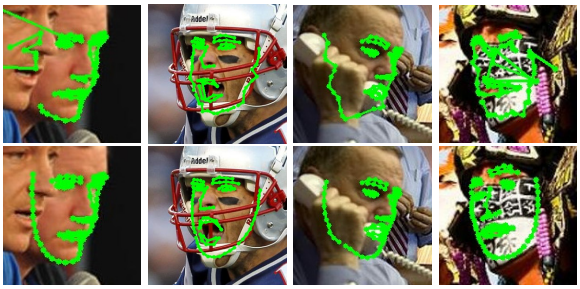


Fig. 2: First and second rows show results obtained using `argmax` over the heatmaps from first ($S = 1$) and second ($S = 2$) stage of our CHR, respectively.

performed on the 300W private competition using the previous 3837 images as training and 600 newly updated images as testing set.

- **COFW**, presented in Burgos-Artizzu et al. (2013), focuses mainly on occlusion. There are 1345 training and 507 testing faces. The annotations include the landmark positions and the binary occlusion labels for 29 points.

- **WFLW** consists of 7500 extremely challenging training and 2500 testing faces divided into six subgroups (pose, expression, illumination, make-up, occlusion and blur), with 98 manually annotated landmarks (Wu et al., 2018).

4.2. Performance metrics

We use common evaluation metrics to quantify 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 constant d_i. Depending on the data base 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

$$NME = \frac{100}{N} \sum_{i=1}^{N} \left(\sum_{l=1}^{L} \left(\frac{\mathbf{w}_i^g(l) \cdot \left\| \mathbf{x}_i(l) - \mathbf{x}_i^g(l) \right\|}{\|\mathbf{w}_i^g\|_1 \cdot d_i} \right) \right). \qquad (4)$$

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.

4.3. Implementation details

To train our algorithms we shuffle each training subset and split it into 90% train and 10% validation. We crop faces using the bounding box annotations enlarged by 30%. In training we also perform data augmentation by applying to each training sample the following random operations: rotation between ±45°, scaling by ±15%, translation by ±5% of the bounding box size, horizontal flip with probability 0.5, colour change multiplying each HSV channel by a random value between [0.5, 1.5] and synthetic rectangular occlusions. See sample results in Fig. 3.

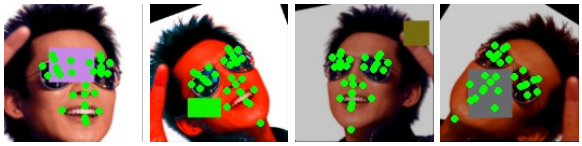


Fig. 3: Data augmentation including random synthetic occlusions.

For training CHR2C 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 0.001. When the validation error levels out for 10 epochs, we multiply the learning rate by 0.5. The fraction of heatmaps removed between stages is set to $f = 0.5$.

We perform experiments using three different configurations: CHR ($S = 1$) and CHR ($S = 2$) using the `argmax` of the heatmaps to compute the 2D landmarks coordinates and CHR2C adding our estimator of coordinates from heatmaps. Training the last configuration end-to-end using WFLW takes 48 hours using a NVidia GeForce GTX 1080Ti GPU (11GB) with a batch size of 6 images.

To ease the reproduction of our results we will release our implementation after publication.

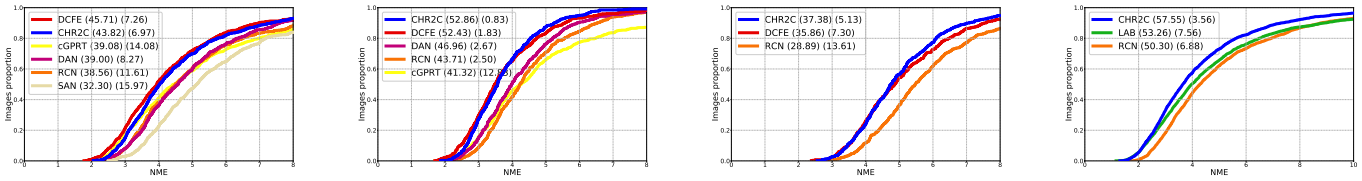
4.4. Experiments with public code

In our first experiment we train SAN (Dong et al., 2018), LAB (Wu et al., 2018), DCFE (Valle et al., 2018), DAN (Kowalski et al., 2017), RCN (Honari et al., 2016) and cGPRT (Lee et al., 2015) with the same settings, including same training, validation and bounding boxes. In Fig. 4 we plot the CED curves. We also provide the AUC_ε and FR_ε values for each method in the legend. We can see that our full approach, CHR2C, reports the highest AUC and smallest FR values in all data sets, except for the public 300W. In this specific case the face shape enforced by DCFE effectively achieves better performance, except for the most difficult faces, with FR_ε above 8, for which the FR of our approach is smaller. In all experiments our CED curve is consistently above the rest, except for the cGPRT and DCFE algorithms "easy" samples, faces with small NME. In the 300W public data set, cGPRT reports better results in samples with NME below 3.0, but our approach is much better in the difficult cases. Most faces in the 300W data set are frontal. This is the reason why the CED curve of DCFE is above ours (see Fig. 4a). This result means that a valid face shape is important to reduce the error when the landmarks are already near their correct location. However, in difficult cases, *i.e.*, 300W public with NME above 8, 300W private with NME above 4, COFW and WFLW, the CED of our approach is above the rest.

4.5. Experiments with published results

In the second set of experiments we compare our method with those in the literature for the 300W challenge (see Tables 1 and 2), COFW (see Table 3) and WFLW (see Table 4).

The 300W public data set has mostly frontal faces. So, an approach such as that in Feng et al. (2018), that introduces a loss conceived to pay more attention to the minimization of samples with small errors, is the one with the best reported result



(a) 300W public (b) 300W private (c) COFW (d) WFLW

Fig. 4: Cumulative error distributions sorted by AUC for each data set.

Table 1: Face alignment results on the 300W public test set.

Method	Common		Challenging		Full			
	pupils	corners	pupils	corners	pupils	corners		
	NME	*NME*	*NME*	*NME*	*NME*	*NME*	AUC_8	FR_8
SAN (Dong et al., 2018)	-	3.34	-	6.60	-	3.98	-	-
cGPRT (Lee et al., 2015)	-	-	-	-	5.71	-	-	-
RCN (Honari et al., 2016)	4.67	-	8.44	-	5.41	-	-	-
ECT (Zhang et al., 2018)	4.66	-	7.96	-	5.31	-	-	-
DAN (Kowalski et al., 2017)	4.42	3.19	7.57	5.24	5.03	3.59	55.33	1.16
TSR (Lv et al., 2017)	4.36	-	7.56	-	4.99	-	-	-
RAR (Xiao et al., 2016)	4.12	-	8.35	-	4.94	-	-	-
RCN^+ (Honari et al., 2018)	4.20	-	7.78	-	4.90	-	-	-
CRN (López et al., 2018)	4.12	2.97	7.90	5.47	4.83	3.44	57.44	1.88
SHN (Yang et al., 2017)	4.12	-	**7.00**	**4.90**	4.68	-	-	-
DU-Net (Tang et al., 2018)	-	2.82	-	5.07	-	3.26	-	-
DCFE (Valle et al., 2018)	3.83	**2.76**	7.54	5.22	4.55	**3.24**	**60.13**	1.59
Wing (Feng et al., 2018)	**3.27**	-	7.18	-	**4.04**	-	-	-
CHR ($S = 1$)	4.21	3.03	8.65	5.99	5.08	3.61	56.28	3.04
CHR ($S = 2$)	4.04	2.91	7.58	5.25	4.73	3.37	58.09	1.45
CHR2C	3.96	2.85	7.44	5.15	4.64	3.30	58.92	**1.16**

Table 2: Face alignment results on the 300W private set.

Method	Indoor	Outdoor	Full				
	corners	corners	corners				
	NME	*NME*	*NME*	AUC_8	FR_8	AUC_{10}	FR_{10}
ECT (Zhang et al., 2018)	-	-	-	45.98	3.17	-	-
DAN (Kowalski et al., 2017)	-	-	4.30	47.00	2.67	-	-
CRN (López et al., 2018)	4.28	4.25	4.26	47.35	2.33	-	-
SHN (Yang et al., 2017)	4.10	4.00	4.05	-	-	-	-
DCFE (Valle et al., 2018)	3.96	3.81	3.88	52.42	1.83	-	-
LAB (Wu et al., 2018)	-	-	-	-	-	58.85	0.83
CHR ($S = 1$)	4.29	4.27	4.28	47.88	3.50	57.85	1.50
CHR ($S = 2$)	3.90	3.89	3.90	51.35	1.00	60.97	0.16
CHR2C	**3.78**	**3.77**	**3.77**	**52.85**	**0.83**	**61.82**	**0.00**

(see Table 1). Here we improve the baseline result reported in López et al. (2018) and other related approaches (Honari et al., 2016, 2018) thanks to the changes introduced in the encoder-decoder architecture, the training procedure and the cascaded configuration. Among the reported results in the private 300W benchmark our full approach is the one with the best performance.

In COFW we report the best result in the literature (see Table 3). This data set has, on average, 28% of the landmarks occluded. However, face poses are mostly frontal. Here, an algorithm like DCFE, that is able to enforce a valid face shape (Valle et al., 2018), provides the second best result. Our cascade, trained taking occlusions into account (CHR ($S = 2$) in Table 3), is on par with DCFE. However, the addition of the final HR2C module (CHR2C in Table 3) is able to improve our estimation of the 2D landmark coordinates establishing a new state-of-the-art in COFW with a NME of 5.09.

Table 3: Face alignment results on COFW.

Method	pupils
	NME
RAR (Xiao et al., 2016)	6.03
ECT (Zhang et al., 2018)	5.98
SHN (Yang et al., 2017)	5.6
LAB (Wu et al., 2018)	5.58
CRN (López et al., 2018)	5.49
Wing (Feng et al., 2018)	5.44
DCFE (Valle et al., 2018)	5.27
CHR ($S = 1$)	6.02
CHR ($S = 2$)	5.30
CHR2C	**5.09**

Similarly in the most challenging data set (as far as we know), WFLW, we achieve the best published results (see Table 4) with an overall NME of 4.39. Note here the large improvement achieved by the second stage of our cascade, that beats by a large margin the previous state-of-the-art in Wu et al. (2018). Moreover, we still get a significant improvement using the HR2C module. This improvement is highest in the *pose* and *occlusion* subsets. The ones with the most challenging images, where the heatmaps become ambiguous.

At run-time our method requires on average 90 ms to process a detected face, a rate of 11 FPS. This processing speed could be halved reducing the number of CNN stages, at the expense of a slight reduction in accuracy (see CHR ($S = 1$)) in Tables 1, 2, 3 and 4).

Finally, in Fig. 5, we report qualitative results for all data sets. In the first three columns we have good alignments and in the last three we have failure cases (NME greater than 8 for 300W, COFW and NME greater than 10 for WFLW). In general, our approach is able to estimate the landmarks always near the ground truth position. As shown in Fig. 2 the cascade of encoder-decoder networks is able to improve the face shape estimation without imposing any parametric model. However, occlusions are still able to drive some landmarks estimations out of the right positions (see mouth on sixth image in Fig. 5a). Also, extreme face poses and expressions can not be perfectly estimated by our method (see fourth image in Fig. 5a, sixth image in Fig. 5b or fourth image in Fig. 5d). Finally, the make-up can be also a problem for our method (see fifth image in Fig. 5d)

Table 4: Face alignment results on WFLW.

Method	Full corners			Pose corners			Expression corners			Illumination corners			Make-up corners			Occlusion corners			Blur corners		
	NME	AUC_{10}	FR_{10}	NME	AUC_{10}	FR_{10}	NME	AUC_{10}	FR_{10}	NME	AUC_{10}	FR_{10}	NME	AUC_{10}	FR_{10}	NME	AUC_{10}	FR_{10}	NME	AUC_{10}	FR_{10}
LAB (Wu et al., 2018)	5.27	53.23	7.56	10.24	23.45	28.83	5.51	49.51	6.37	5.23	54.33	6.73	5.15	53.94	7.77	6.79	44.90	13.72	6.32	46.30	10.74
CHR ($S = 1$)	5.02	53.48	6.64	9.33	24.89	26.99	5.44	49.95	6.68	6.68	54.30	5.01	4.91	54.04	7.28	6.41	43.94	13.72	5.75	46.96	8.53
CHR ($S = 2$)	4.57	56.39	4.20	8.10	29.70	20.24	4.89	53.55	3.50	4.58	56.98	3.29	4.36	57.24	3.39	5.64	48.00	8.28	5.28	50.24	6.20
CHR2C	**4.39**	**57.55**	**3.55**	**7.58**	**31.85**	**18.09**	**4.72**	**55.04**	**3.82**	**4.39**	**57.94**	**2.57**	**4.18**	**58.82**	**1.94**	**5.37**	**49.63**	**7.06**	**5.09**	**51.54**	**5.30**

as there are not many training faces with it.

5. Conclusions

In this paper we have introduced CHR2C, a facial landmark detection algorithm that exploits the benefits of a high capacity cascade of CNN regressors. As shown in the experiments, this additional capacity is crucial to improve the estimated landmark location in difficult poses and occlusions.

In our approach we improve the regressor by cascading two identical networks and training it in a way that takes occlusions into account. We additionally add a simple but important final stage, termed HR2C, that improves by a significant margin the usual `argmax` approach to estimate landmark coordinates from heatmaps. We have shown in the experiments that it notably improves the results in WFLW and lets us establish a new state-of-the-art in COFW.

An alternative approach, DCFE (Valle et al., 2018), aims to enforce a valid face shape on the set of landmark heatmaps. This improves the final accuracy when the heatmaps provide a good approximate estimation for the location of all landmarks. However, in challenging situations, with extreme poses and occlusions, improving the regressor capacity is more important as we have shown in the experiments.

In a future work we plan to study the combination of high capacity regressors with the enforcement face shape, to get the best of both approaches.

Acknowledgments

The authors acknowledge funding from the Spanish Ministry of Economy and Competitiveness under project TIN2016-75982-C2-2-R. They are also grateful to Pedro D. López for programming the first version of this model.

References

Bekios-Calfa, J., Buenaposada, J.M., Baumela, L., 2014. Robust gender recognition by exploiting facial attributes dependencies. Pattern Recognition Letters 36, 228–234.

Bhattacharya, S., Nainala, G.S., Rooj, S., Routray, A., 2019. Local force pattern (LFP): Descriptor for heterogeneous face recognition. Pattern Recognition Letters 125, 63–70.

Burgos-Artizzu, X.P., Perona, P., Dollar, P., 2013. Robust face landmark estimation under occlusion, in: Proc. International Conference on Computer Vision, pp. 1513–1520.

Dollar, P., Welinder, P., Perona, P., 2010. Cascaded pose regression, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1078–1085.

Dong, X., Yan, Y., Ouyang, W., Yang, Y., 2018. Style aggregated network for facial landmark detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 379–388.

Feng, Z., Kittler, J., Awais, M., Huber, P., Wu, X., 2018. Wing loss for robust facial landmark localisation with convolutional neural networks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245.

Guo, J., Deng, J., Xue, N., Zafeiriou, S., 2018. Stacked dense u-nets with dual transformers for robust face alignment, in: Proc. British Machine Vision Conference, p. 44.

Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C.J., Kautz, J., 2018. Improving landmark localization with semi-supervised learning, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1546–1555.

Honari, S., Yosinski, J., Vincent, P., Pal, C.J., 2016. Recombinator networks: Learning coarse-to-fine feature aggregation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5743–5752.

Kowalski, M., Naruniec, J., Trzcinski, T., 2017. Deep alignment network: A convolutional neural network for robust face alignment, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2034–2043.

Lee, D., Park, H., Yoo, C.D., 2015. Face alignment using cascade gaussian process regression trees, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4204–4212.

López, P.D., Valle, R., Baumela, L., 2018. Facial landmarks detection using a cascade of recombinator networks, in: Proc. Iberoamerican Congress on Pattern Recognition, pp. 575–583.

Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X., 2017. 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, pp. 3691–3700.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation, in: Proc. European Conference on Computer Vision, pp. 483–499.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2016. 300 faces in-the-wild challenge: database and results. Image and Vision Computing 47, 3–18.

Sun, N., Li, Q., Huan, R., Liu, J., Han, G., 2019. Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recognition Letters 119, 49–61.

Tang, Z., Peng, X., Geng, S., Wu, L., Zhang, S., Metaxas, D.N., 2018. Quantized densely connected u-nets for efficient landmark localization, in: Proc. European Conference on Computer Vision, pp. 348–364.

Valle, R., Buenaposada, J.M., Valdés, A., Baumela, L., 2018. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment, in: Proc. European Conference on Computer Vision, pp. 609–624.

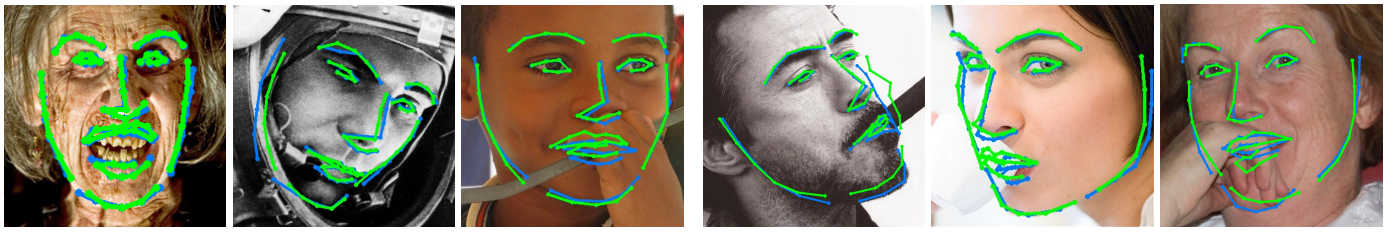
Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q., 2018. Look at boundary: A boundary-aware face alignment algorithm, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2138.

Wu, Y., Ji, Q., 2019. Facial landmark detection: A literature survey. International Journal of Computer Vision 127, 115–142.

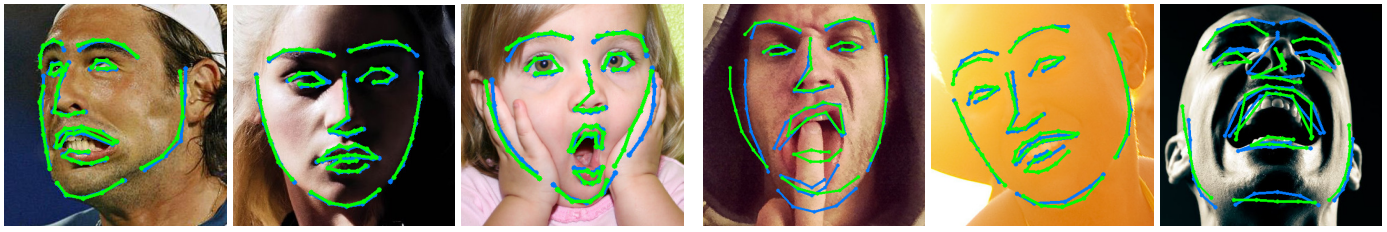
Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.A., 2016. Robust facial landmark detection via recurrent attentive-refinement networks, in: Proc. European Conference on Computer Vision, pp. 57–72.

Yang, J., Liu, Q., Zhang, K., 2017. Stacked hourglass network for robust facial landmark localisation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2025–2033.

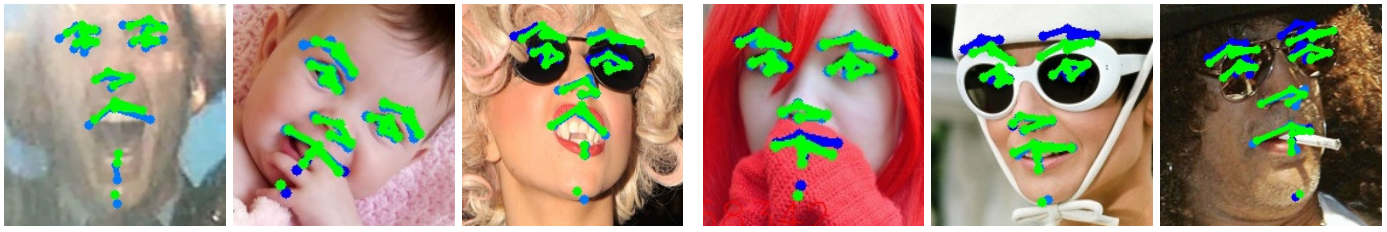
Zhang, H., Li, Q., Sun, Z., Liu, Y., 2018. Combining data-driven and model-driven methods for robust facial landmark detection. IEEE Trans. Information Forensics and Security 13, 2409–2422.



(a) 300W public



(b) 300W private



(c) COFW

(d) WFLW

Fig. 5: Representative results using CHR2C in 300W public/private, COFW and WFLW testing subsets. Blue and green colors represent ground truth and shape predictions respectively. The first three columns show successful face alignment and three last columns some of the worst results according to FR_{ε}.