A Deeply-initialized Coarse-to-fine Ensemble of Regression Trees for Face Alignment

Roberto Valle$^{1[0000-0003-1423-1478]}$, José M. Buenaposada$^{2[0000-0002-4308-9653]}$, Antonio Valdés^{3}, and Luis Baumela^{1}

[1] Univ. Politécnica de Madrid, Spain. {rvalle,lbaumela}@fi.upm.es
[2] Univ. Rey Juan Carlos, Spain. josemiguel.buenaposada@urjc.es
[3] Univ. Complutense de Madrid, Spain. avaldes@ucm.es

Abstract. In this paper we present DCFE, a real-time facial landmark regression method based on a coarse-to-fine Ensemble of Regression Trees (ERT). We use a simple Convolutional Neural Network (CNN) to generate probability maps of landmarks location. These are further refined with the ERT regressor, which is initialized by fitting a 3D face model to the landmark maps. The coarse-to-fine structure of the ERT lets us address the combinatorial explosion of parts deformation. With the 3D model we also tackle other key problems such as robust regressor initialization, self occlusions, and simultaneous frontal and profile face analysis. In the experiments DCFE achieves the best reported result in AFLW, COFW, and 300W private and common public data sets.

Keywords: Face alignment, Cascaded Shape Regression, Convolutional Neural Networks, Coarse-to-fine, Occlusions, Real-time

1 Introduction

Facial landmarks detection is a preliminary step for many face image analysis problems such as verification and recognition [25], attributes estimation [2], etc. The availability of large annotated data sets has recently encouraged research in this area with important performance improvements. However, it is still a challenging task especially when the faces suffer from large pose variations and partial occlusions.

The top performers in the recent 300W benchmark are all based in deep regression models [20, 23, 30, 33] (see Table 1). The most prominent feature of these approaches is their robustness, due to the large receptive fields of deep nets. However, in these models it is not easy to enforce facial shape consistency or estimate self occlusions.

ERT-based models [6, 7, 18, 24], on the other hand, are easy to parallelize and implicitly impose shape consistency in their estimations. They are much more efficient than deep models and, as we demonstrate in our experiments (see Fig. 4), with a good initialization, they are also very accurate.

In this paper we present a hybrid method, termed Deeply-initialized Coarse-to-Fine Ensemble (DCFE). It uses a simple CNN to generate probability maps of landmarks location. Hence, obtaining information about the position of individual landmarks without a globally imposed shape. Then we fit a 3D face model, thus enforcing a global face shape prior. This is the starting point of the coarse-to-fine ERT regressor. The fitted 3D face model provides the regressor with a valid initial shape and information about landmarks visibility. The coarse-to-fine approach lets the ERT easily address the combinatorial explosion of all possible deformations of non-rigid parts and at the same time impose a part shape prior. The proposed method runs in real-time (32 FPS) and provides the best reported results in AFLW, COFW, and 300W private and common public data sets.

2 Related Work

Face alignment has been a topic of intense research for more than twenty years. Initial successful results were based on 2D and 3D generative approaches such as the Active Appearance Models (AAM) [8] or the 3D Morphable Models [4]. More recent discriminative methods are based on two key ideas: indexing image description relative to the current shape estimate [12] and the use of a regressor whose predictions lie on the subspace spanned by the training face shapes [7], this is the so-called Cascade Shape Regressor (CSR) framework. Kazemi *et al.* [18] improved the original cascade framework by proposing a real-time ensemble of regression trees. Ren *et al.* [24] used locally binary features to boost the performance up to 3000 FPS. Burgos-Artizzu *et al.* [6] included occlusion information. Xiong *et al.* [31, 32] use SIFT features and learn a linear regressor dividing the search space into individual regions with similar gradient directions. Overall, the CSR approach is very sensitive to the starting point of the regression process. An important part of recent work revolves around how to find good initialisations [38, 37]. In this paper we use the landmark probability maps produced by a CNN to find a robust starting point for the CSR.

Current state-of-the-art methods in face alignment are based on CNNs. Sun *et al.* [26] were pioneers to apply a three-level CNN to obtain accurate landmark estimation. Zhang *et al.* [36] proposed a multi-task solution to deal with face alignment and attributes classification. Lv *et al.*'s [23] uses global and local face parts regressors for fine-grained facial deformation estimation. Yu *et al.* [34] transforms the landmarks rather than the input image for the refinement cascade. Trigeorgis *et al.* [27] and Xiao *et al.* [30] are the first approaches that fuse the feature extraction and regression steps of CSR into a recurrent neural network trained end-to-end. Kowalski *et al.* [20] and Yang *et al.* [33] are among the top performers in the Menpo competition [35]. Both 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. The large receptive fields of deep neural nets convey these approaches with a high degree of robustness to face rotation, scale and deformation. However, it is not clear how to impose facial shape consistency on the estimated set of landmarks. Moreover, to achieve accuracy they resort to a cascade of deep models that progressively refine the estimation, thus incrementing the computational requirements.

There is also an increasing number of works based on 3D face models. In the simplest case they fit a mean model to the estimated image landmarks position [19] or jointly regress the pose and shape of the face [17, 29]. These approaches provide 3D pose information that may be used to estimate landmark self-occlusions or to train simpler regressors specialized in a given head orientation. However, building and fitting a 3D face model is a difficult task and the results of the full 3D approaches in current benchmarks are not as good as those described above.

Our proposal tries to leverage the best features of the previous approaches. Using a CCN-based initialization we inherit the robustness of deep models. Like the simple 3D approaches we fit a rigid 3D face model to initialize the ERT and estimate global face orientation to address self occlusions. Finally, we use an ERT within a coarse-to-fine framework to achieve accuracy and efficiency.

3 Deeply Initialized Coarse-to-Fine Ensemble

In this section, we present the Deeply-initialized Coarse-to-fine Ensemble method (DCFE). It consists of two main steps: CNN-based rigid face pose computation and ERT-based non-rigid face deformation estimation, both shown in Fig. 1.

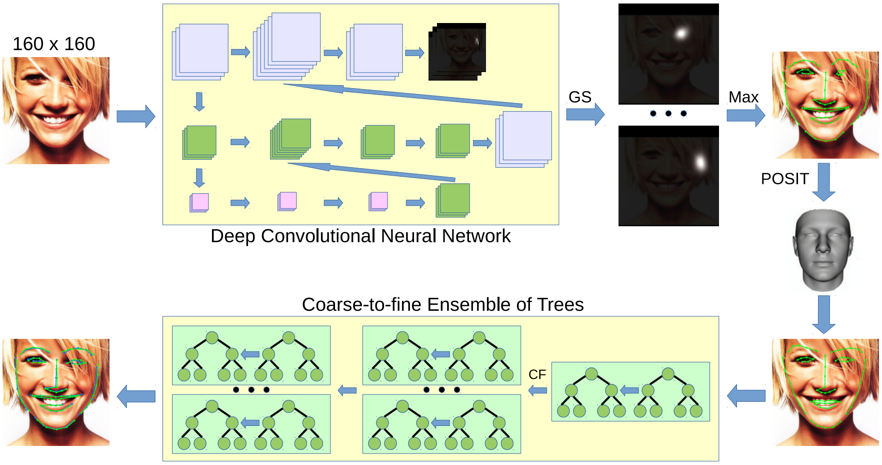

160 x 160
Deep Convolutional Neural Network
GS
Max
POSIT
Coarse-to-fine Ensemble of Trees
CF

Fig. 1: DCFE framework diagram. GS, Max and POSIT represent the Gaussian smoothing filter, the maximum of each probability map and the 3D pose estimation respectively.

3.1 Rigid pose computation

ERT-based regressors require an acceptable initialization to converge to a good solution. We propose the use of face landmark location probability maps like [3, 9, 30] to generate plausible shape initialization candidates. We have modified Honari *et al.*'s [16] RCN introducing a loss function to handle missing landmarks, thus enabling semi-supervised training. We train this CNN to obtain a set of probability maps, $\mathcal{P}(\mathtt{I})$, indicating the position of each landmark in the input image (see Fig. 1). The maximum of each smoothed probability map determines our initial landmark positions. Note in Fig. 1 that these predictions are sensitive to occlusions and may not be a valid face shape. Compared to typical CNN-based approaches, *e.g.*, [33], our CNN is simpler, since we only require a rough estimation of landmark locations.

To start the ERT with a plausible face, we compute the initial shape by fitting a rigid 3D head model to the estimated 2D landmarks locations. To this end we use the softPOSIT algorithm proposed by David *et al.* [10]. As a result, we project the 3D model onto the image using the estimated rigid transformation. This provides the ERT with a rough estimation of the scale, translation and 3D pose of the target face (see Fig. 1).

Let $\mathbf{x}^0 = g_0(\mathcal{P}(\mathtt{I}))$ be the *initial shape*, the output of the initialization function g_0 after processing the input image $\mathtt{I}$. In this case $\mathbf{x}^0$ is a $L \times 2$ vector with L 2D landmarks coordinates. With our initialization we ensure that $\mathbf{x}^0$ is a valid face shape. This guarantees that the predictions in the next step of the algorithm will also be valid face shapes [7].

3.2 ERT-based non-rigid shape estimation

Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ be the set of train face shapes, where $s_i = (\mathtt{I}_i, \mathbf{x}_i^g, \mathbf{v}_i^g, \mathbf{w}_i^g, \mathbf{x}_i^0)$. Each training shape s_i has its own: training image, $\mathtt{I}_i$; ground truth shape, $\mathbf{x}_i^g$; ground truth visibility label, $\mathbf{v}_i^g$; annotated landmark label, $\mathbf{w}_i^g$ (1 annotated and 0 missing) and initial shape for regression training, $\mathbf{x}_i^0$. The ground truth (or target) shape, $\mathbf{x}_i^g$, is a $L \times 2$ vector with the L landmarks coordinates. The $L \times 1$ vector $\mathbf{v}_i^g$ holds the visibility binary label of each landmark. If the k-th component of $\mathbf{v}^g$, $\mathbf{v}^g(k) = 1$ then the k-th landmark is visible. In our implementation we use shape-indexed features [21], $\phi(\mathcal{P}(\mathtt{I}_i), \mathbf{x}_i^t, \mathbf{w}_i^g)$, that depend on the current shape $\mathbf{x}_i^t$ of the landmarks in image $\mathtt{I}_i$ and whether they are annotated or not, $\mathbf{w}_i^t$.

We divide the regression process into T stages and learn an ensemble of K regression trees for the t-th stage, $\mathcal{C}_t(f_i) = \mathbf{x}^{t-1} + \sum_{k=1}^{K} g_k(f_i)$, where $f_i = \phi(\mathcal{P}(\mathtt{I}), \mathbf{x}^{t-1}, \mathbf{w}_i^g)$ and $\mathbf{x}^j$ are the coordinates of the landmarks estimated in j-th stage (or the initialization coordinates, $\mathbf{x}^0$, in the first stage). To train the whole ERT we use the N training samples in $\mathcal{S}$ to generate an augmented training set, S_A with cardinality $N_A = |\mathcal{S}_A|$. From each training shape s_i we generate additional training samples by changing their initial shape. To this end we randomly sample new candidate landmark positions from the smoothed probability maps to generate the new initial shapes (see section 3.1).

We incorporate the visibility label $\mathbf{v}$ with the shape to better handle occlusions (see Fig. 5c) in a way similar to Burgos-Artizzu *et al.* [6] and naturally handling partially labelled training data like Kazemi *et al.* [18] using ground-truth annotation labels $\mathbf{w} \in \{0, 1\}$. Each initial shape is progressively refined by estimating a shape and visibility increments $\mathcal{C}_t^{\mathbf{v}}(\phi(\mathcal{P}(\mathtt{I}_i), \mathbf{x}_i^{t-1}, \mathbf{w}_i^g))$ where $\mathbf{x}_i^{t-1}$ represents the current shape of the i-th sample (see Algorithm 1). $\mathcal{C}_t^{\mathbf{v}}$ is trained to minimize only the landmark position errors but on each tree leaf, in addition to the mean shape, we also output the mean of all training shapes visibilities, $\mathbf{v}_i^g$, that belong to that node. We define $\mathcal{U}_{t-1} = \{(\mathbf{x}_i^{t-1}, \mathbf{v}_i^{t-1})\}_{i=1}^{N_A}$ as the set of all current shapes and corresponding visibility vectors for all training data.

Algorithm 1 Training an Ensemble of Regression Trees

```
Input: Training data $\mathcal{S}$, $T$
  Generate augmented training samples set, $\mathcal{S}_A$
  for t=1 to $T$ do
    Extract features for all samples, $\mathcal{F}_A = \{f_i\}_{i=1}^{N_A} = \{\phi(\mathcal{P}(\mathtt{I}_i), \mathbf{x}_i^{t-1}, \mathbf{w}_i^g)\}_{i=1}^{N_A}$
    Learn coarse-to-fine regressor, $\mathcal{C}_t^{\mathbf{v}}$, from $\mathcal{S}_\mathcal{A}$, $\mathcal{F}_A$ and $\mathcal{U}_{t-1} = \{(\mathbf{x}_i^{t-1}, \mathbf{v}_i^{t-1})\}_{i=1}^{N_A}$
    Update current shapes and visibilities, $\{(\mathbf{x}_i^t, \mathbf{v}_i^t) = (\mathbf{x}_i^{t-1}, \mathbf{v}_i^{t-1}) + \mathcal{C}_t^{\mathbf{v}}(f_i)\}_{i=1}^{N_A}$
  end for
Output: $\{\mathcal{C}_t^{\mathbf{v}}\}_{t=1}^{T}$
```

Compared with conventional ERT approaches, our ensemble is simpler. It will require fewer trees because we only have to estimate the non-rigid face shape deformation, since the 3D rigid component has been estimated in the previous step. In the following, we describe the details of our ERT.

Initial shapes for regression. The selection of the starting point in the ERT is fundamental to reach a good solution. The simplest choice is the mean of the ground truth training shapes, $\bar{x}^0 = \sum_{i=1}^{N} \mathbf{x}_i^g / N$. However, such a poor initialization leads to wrong alignment results in test images with large pose variations. Alternative strategies are running the ERT several times with different initializations and taking the median [6], initializing with other ground truth shapes $x_i^0 \leftarrow x_j^g$ where $i \neq j$ [18] or randomly deforming the initial shape [20].

In our approach we initialize the ERT using the algorithm described in section 3.1, that provides a robust and approximate shape for initialization (see Fig. 2). Hence, the ERT only needs to estimate the non-rigid component of face pose.

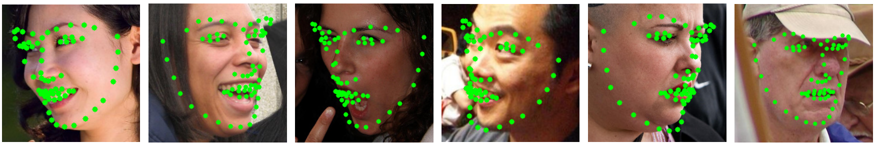


Fig. 2: Worst initial shapes for the 300W training subset.

Feature Extraction. ERT efficiency depends on the feature extraction step. In general, descriptor features such as SIFT used by [31, 38] improve face alignment results, but have higher computational cost compared to simpler features such as plain pixel value differences [7, 6, 18, 24]. In our case, a simple feature suffices, since shape landmarks are close to their ground truth location.

In DCFE we use the probability maps $\mathcal{P}(\mathtt{I})$ to extract features for the cascade. To this end, we select a landmark l and its associated probability map $\mathcal{P}^l(\mathtt{I})$. The feature is computed as the difference between two pixels values in $\mathcal{P}^l(\mathtt{I})$ from a FREAK descriptor pattern [1] around l. Our features are similar to those in Lee *et al.*'s [21]. However, ours are defined on the probability maps, $\mathcal{P}(\mathtt{I})$, instead of the image, $\mathtt{I}$. We let the training algorithm select the most informative landmark and pair of pixels in each iteration.

Learn a coarse-to-fine regressor. To train the t-th stage regressor, $\mathcal{C}_t^{\mathbf{v}}$, we fit an ERT. Thus, the goal is to sequentially learn a series of weak learners to greedily minimize the regression loss function:

$$\mathcal{L}_t(\mathcal{S}_A, \mathcal{F}_A, \mathcal{U}_{t-1}) = \sum_{i=1}^{N_A} ||\mathbf{w}_i^g \odot (\mathbf{x}_i^g - \mathbf{x}_i^{t-1} - \sum_{k=1}^{K} g_k(f_i))||^2, \tag{1}$$

where $\odot$ is the Hadamard product. There are different ways of minimizing Equation 1. Kazemi *et al.* [18] present a general framework based on Gradient Boosting for learning an ensemble of regression trees. Lee *et al.* [21] establish an optimization method based on Gaussian Processes also learning an ensemble of regression trees but outperforming previous literature by reducing the overfitting.

A crucial problem when training a global face landmark regressor is the lack of examples showing all possible combinations of face parts deformations. Hence, these regressors quickly overfit and generalize poorly to combinations of part deformations not present in the training set. To address this problem we introduce the coarse-to-fine ERT architecture.

The goal is to be able to cope with combinations of face part deformations not seen during training. A single monolithic regressor is not able to estimate these local deformations (see difference between Fig. 3b and Fig. 3c). Our algorithm is agnostic in the number of parts or levels of the coarse-to-fine estimation. Algorithm 2 details the training of P face parts regressors (each one with a subset of the landmarks) to build a coarse-to-fine regressor. Note that $\mathbf{x}_i^0$ and $\mathbf{v}_i^0$ in this context are the shape and visibility vectors from the last regressor output (*e.g.*, the previous part regressor or a previous full stage regressor). In our implementation we use $P = 1$ (all landmarks) with the first K_1 regressors and in the last K_2 regressors the number of parts is increased to $P = 10$ (left/right eyebrow, left/right eye, nose, top/bottom mouth, left/right ear and chin), see all the parts connected by lines in Fig. 3c.

Fit a regression tree. The training objective for the k-th regression tree is to minimize the sum of squared residuals, taking into account the annotated

Algorithm 2 Training P parts regressors

```
Input: $\mathcal{S}_A, \mathcal{F}_A, \{(\mathbf{x}_i^0, \mathbf{v}_i^0)\}_{i=1}^{N_A}, \nu, K, P$
  for k=1 to $K$ do
    for p=1 to $P$ do
      // $\odot$ is the Hadamard product, $(p)$ selects elements of a vector in that part
      Compute shape residuals $\{\mathbf{r}_i^k(p) = \mathbf{w}_i^g(p) \odot (\mathbf{x}_i^g(p) - \mathbf{x}_i^{k-1}(p)\}_{i=1}^{N_A}$
      Fit a regression tree $g_k^p$ using the residuals $\{\mathbf{r}_i^k(p)\}$ and $\mathcal{F}_A(p)$
      // $\nu$ is the shrinkage factor to scale the contribution of each tree
      Update samples $\{(\mathbf{x}_i^k(p), \mathbf{v}_i^k(p)) = (\mathbf{x}_i^{k-1}(p), \mathbf{v}_i^{k-1}(p)) + \nu \cdot g_k^p(f_i(p))\}_{i=1}^{N_A}$
    end for
  end for
Output: $P$ part regressors $\{\mathcal{C}^p\}_{p=1}^P$, with $K$ weak learners each $\mathcal{C}^p = \{g_k^p\}_{k=1}^K$
```

landmark labels:

$$\mathcal{E}_k = \sum_{i=1}^{N_A} ||\mathbf{r}_i^k||^2 = \sum_{i=1}^{N_A} ||\mathbf{w}_i^g \odot (\mathbf{x}_i^g - \mathbf{x}_i^{k-1})||^2. \tag{2}$$

We learn each regression binary tree by recursively splitting the training set into the left (l) and right (r) child nodes. The tree node split function is designed to minimize $\mathcal{E}_k$ from Equation 2 in the selected landmark. To train a regression tree node we randomly generate a set of candidate split functions, each of them involving four parameters $\theta = (\tau, \mathbf{p}_1, \mathbf{p}_2, l)$, where $\mathbf{p}_1$ and $\mathbf{p}_2$ are pixels coordinates on a fixed FREAK structure around the l-th landmark coordinates in $\mathbf{x}_i^{k-1}$. The feature value corresponding to θ for the i-th training sample is $f_i(\theta) = \mathcal{P}^l(\mathtt{I}_i)[\mathbf{p}_1] - \mathcal{P}^l(\mathtt{I}_i)[\mathbf{p}_2]$, the difference of probability values in the maps for the given landmark. Finally, we compute the split function thresholding the feature value, $f_i(\theta) > \tau$.

Given $\mathcal{N} \subset \mathcal{S}_A$ the set of training samples at a node, fitting a tree node for the k-th tree, consists of finding the parameter θ that minimizes $E_k(\mathcal{N}, \theta)$

$$\arg\min_\theta E_k(\mathcal{N}, \theta) = \arg\min_\theta \sum_{b \in \{l,r\}} \sum_{s \in \mathcal{N}_{\theta,b}} ||\mathbf{r}_s^k - \boldsymbol{\mu}_{\theta,b}||^2 \tag{3}$$

where $\mathcal{N}_{\theta,l}$ and $\mathcal{N}_{\theta,r}$ are, respectively, the samples sent to the left and right child nodes due to the decision induced by θ. The mean residual $\boldsymbol{\mu}_{\theta,b}$ for a candidate split function and a subset of training data is given by

$$\boldsymbol{\mu}_{\theta,b} = \frac{1}{|\mathcal{N}_{\theta,b}|} \sum_{s \in \mathcal{N}_{\theta,b}} \mathbf{r}_s^k \tag{4}$$

Once we know the optimal split each leaf node stores the mean residual, $\boldsymbol{\mu}_{\theta,b}$, as the output of the regression for any example reaching that leaf.

4 Experiments

To train and evaluate our proposal, we perform experiments with 300W, COFW and AFLW that are considered the most challenging public data sets. In addition, we also show qualitative face alignment results with the Menpo competition images.

- **300W**. It provides bounding boxes and 68 manually annotated landmarks. We follow the most established approach and divide the 300W annotations into 3148 training and 689 testing images (public competition). Evaluation is also performed on the newly updated 300W private competition.
- **Menpo**. Consist of 8979 training and 16259 testing faces containing 12006 semi-frontal and 4253 profile images. The images were annotated with the previous set of 68 landmarks but without facial bounding boxes.
- **COFW**. It focuses on occlusion. Commonly, there are 1345 training faces in total. The testing set is made of 507 images. The annotations include the landmark positions and the binary occlusion labels for 29 points.
- **AFLW**. Provides an extensive collection of 25993 in-the-wild faces, with 21 facial landmarks annotated depending on their visibility. We have found several annotations errors and, consequently, removed these faces from our experiments. From the remaining faces we randomly choose 19312 images for training/validation and 4828 instances for testing.

4.1 Evaluation

We use the Normalized Mean Error (NME) as a metric to measure the shape estimation error

$$NME = \frac{100}{N} \sum_{i=1}^{N} \left(\frac{1}{||\mathbf{w}_i^g||_1} \sum_{l=1}^{L} \left(\frac{\mathbf{w}_i^g(l) \cdot \|\mathbf{x}_i(l) - \mathbf{x}_i^g(l)\|}{d_i} \right) \right). \tag{5}$$

It computes the euclidean distance between the ground-truth and estimated landmark positions normalized by d_i. We report our results using different values of d_i: the distance between the eye centres (*pupils*), the distance between the outer eye corners (*corners*) and the bounding box size (*height*).

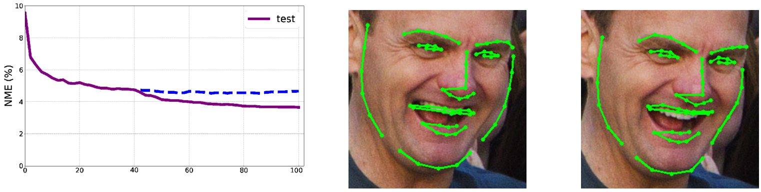
In addition, we also compare our results using Cumulative Error Distribution (CED) curves. We calculate AUC_ε as the area under the CED curve for images with an NME smaller than ε and FR_ε as the failure rate representing the percentage of testing faces with NME greater than ε. We use precision/recall percentages to compare occlusion prediction.

To train our algorithm we shuffle the training set and split it into 90% train-set and 10% validation-set.

4.2 Implementation

All experiments have been carried out with the settings described in this section. We train from scratch the CNN selecting the model parameters with lowest validation error. We crop faces using the original bounding boxes annotations enlarged by 30%. We generate different training samples in each epoch by applying random in plane rotations between ±30°, scale changes by ±15% and translations by ±5% of bounding box size, randomly mirroring images horizontally and generating random rectangular occlusions. We use Adam stochastic optimization with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e^{-8}$ parameters. We train during 400 epochs with an initial learning rate $\alpha = 0.001$, without decay and a batch size of 35 images. In the CNN the cropped input face is reduced from 160×160 to 1×1 pixels gradually dividing by half their size across $B = 8$ branches applying a 2×2 pooling[4]. All layers contain 64 channels to describe the required landmark features.

We train the coarse-to-fine ERT with the Gradient Boosting algorithm [15]. It requires $T = 20$ stages of $K = 50$ regression trees per stage. The depth of trees is set to 5. The number of tests to choose the best split parameters, θ, is set to 200. We resize each image to set the face size to 160 pixels. For feature extraction, the FREAK pattern diameter is reduced gradually in each stage (*i.e.*, in the last stages the pixel pairs for each feature are closer). We generate several initializations for each face training image to create a set of at least $N_A = 60000$ samples to train the cascade. To avoid overfitting we use a shrinkage factor $\nu = 0.1$ in the ERT. Our regressor triggers the coarse-to-fine strategy once the cascade has gone through 40% of the stages (see Fig. 3a).



(a) NME evolution (b) Monolithic (c) Coarse-to-fine

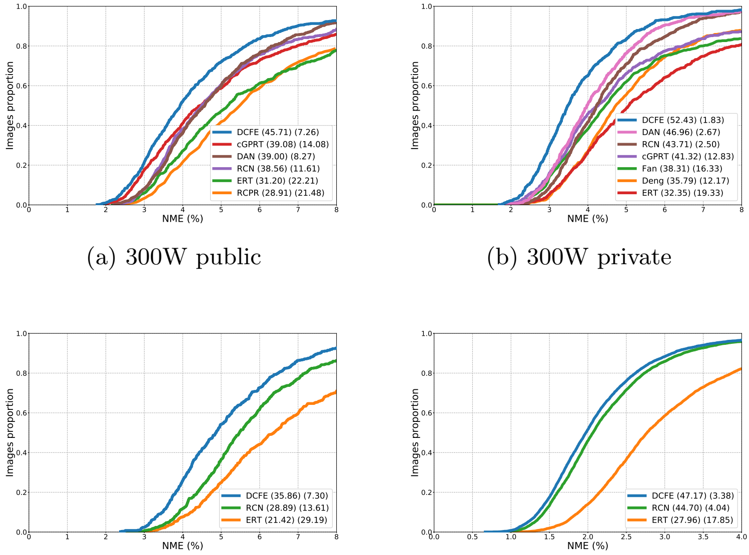
Fig. 3: Example of a monolithic ERT regressor vs our coarse-to-fine approach. (a) Evolution of the error through the different stages in the cascade (dashed line represents the algorithm without the coarse-to-fine improvement); (b) predicted shape with a monolithic regressor; (c) predicted shape with our coarse-to-fine approach.

For the Mempo data set training the CNN and the coarse-to-fine ensemble of trees takes 48 hours using a NVidia GeForce GTX 1080 (8GB) GPU and an Intel Xeon E5-1650 at 3.50GHz (6 cores/12 threads, 32 GB of RAM). At runtime our method process test images on average at a rate of 32 FPS, where the CNN takes 25 ms and the ERT 6.25 ms per face image using C++, Tensorflow and OpenCV libraries.

[4] Except when the 5×5 images are reduced to 2×2 where we apply a 3×3 pooling

4.3 Results

Here we compare our algorithm, DCFE, with the best reported results for each data set. To this end we have trained our model and those in DAN [20], RCN [16], cGPRT [21], RCPR [6] and ERT [18] with the code provided by the authors and the same settings including same training, validation and bounding boxes. In Fig. 4 we plot the CED curves and we provide AUC_8 and FR_8 values for each algorithm. Also, for comparison with other methods in Tables 1, 2, 3, 4 we show the original results published in the literature.



(a) 300W public

(b) 300W private

(c) COFW

(d) AFLW

Fig. 4: Cumulative error distributions sorted by AUC.

In Tables 1 and 2 we provide the results of the state-of-the-art methods in the 300W public and private data sets. Our approach obtains the best performance in the private (see Table 2) and in the common and full subsets of the 300W competition public test set (see Table 1). This is due to the excellent accuracy achieved by the coarse-to-fine ERT scheme enforcing valid face shapes. In the challenging subset of the 300W competition public test set SHN [33] achieves better results. This is caused by errors in initializing the ERT in a few images with very large scale and pose variations, that are not present in the training set. Our method exhibits superior capability in handling cases with low error since we achieve the best NME results in the 300W common subset by the largest margin. The CED curves in Figs. 4a and 4b show that DCFE is better than all its competitors that provide code in all types of images in both data sets. In the 300W private challenge we obtain the best results outperforming Deng *et al.* [11] and Fan *et al.* [13] that were the academia and industry winners of the competition (see Fig. 4b).

Method	Common		Challenging		Full			
	pupils	corners	pupils	corners	pupils	corners		
	NME	NME	NME	NME	NME	NME	AUC_8	FR_8
RCPR [6]	6.18	-	17.26	-	8.35	-	-	-
ESR [7]	5.28	-	17.00	-	7.58	-	43.12	10.45
SDM [31]	5.60	-	15.40	-	7.52	-	42.94	10.89
ERT [18]	-	-	-	-	6.40	-	-	-
LBF [24]	4.95	-	11.98	-	6.32	-	-	-
cGPRT [21]	-	-	-	-	5.71	-	-	-
CFSS [38]	4.73	-	9.98	-	5.76	-	49.87	5.08
DDN [34]	-	-	-	-	5.65	-	-	-
TCDCN [36]	4.80	-	8.60	-	5.54	-	-	-
MDM [27]	-	-	-	-	-	-	52.12	4.21
RCN [16]	4.67	-	8.44	-	5.41	-	-	-
DAN [20]	4.42	3.19	7.57	5.24	5.03	3.59	55.33	1.16
TSR [23]	4.36	-	7.56	-	4.99	-	-	-
RAR [30]	4.12	-	8.35	-	4.94	-	-	-
SHN [33]	4.12	-	7.00	4.90	-	-	-	-
DCFE	3.83	2.76	7.54	5.22	4.55	3.24	60.13	1.59

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

We may appreciate the improvement achieved by the ERT by comparing the results of DCFE in the full subset of 300W, 4.55, with Honari's baseline RCN [16], 5.41. It represents an 16% improvement. The coarse-to-fine strategy in our ERT only affects difficult cases, with rare facial part combinations. Zooming-in Figs. 3b and 3c you may appreciate how it improves the adjustment of the cheek and mouth. Although it is a crucial step to align local parts properly, the global NME is only marginally affected.

Table 3 and Fig. 4c compare the performance of our model and baselines using the COFW data set. We obtain the best results (*i.e.*, NME 5.27) establishing a new state-of-the-art without requiring a sophisticated network, which demonstrates the importance of preserving the facial shape and the robustness of our framework to severe occlusions. In terms of landmark visibility, we have obtained comparable performance with previous methods.

In Table 4 and Fig. 4d we show the results with AFLW. This is a challenging data set not only because of its size, but also because of the number of samples with self-occluded landmarks that are not annotated. This is the reason for the small number of competitors in Fig. 4d, very few approaches allow training with

Method	Indoor corners			Outdoor corners			Full corners		
	NME	AUC_8	FR_8	NME	AUC_8	FR_8	NME	AUC_8	FR_8
ESR [7]	-	-	-	-	-	-	-	32.35	17.00
cGPRT [21]	-	-	-	-	-	-	-	41.32	12.83
CFSS [38]	-	-	-	-	-	-	-	39.81	12.30
MDM [27]	-	-	-	-	-	-	5.05	45.32	6.80
DAN [20]	-	-	-	-	-	-	4.30	47.00	2.67
SHN [33]	4.10	-	-	4.00	-	-	4.05	-	-
DCFE	3.96	52.28	2.33	3.81	52.56	1.33	3.88	52.42	1.83

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

missing data. Although the results in Table 4 are not strictly comparable because each paper uses its own train and test subsets, we get a NME of 2.17 that again establishes a new state-of-the-art, considering that [37, 14, 23] do not use the two most difficult landmarks, the ones in the ears.

Method	pupils			occlusion
	NME	AUC_8	FR_8	precision/recall
ESR [7]	11.20	-	-	-
RCPR [6]	8.50	-	-	80/40
TCDCN [36]	8.05	-	-	-
RAR [30]	6.03	-	-	-
DAC-CSR [14]	6.03	-	-	-
Wu *et al.* [28]	5.93	-	-	80/49.11
SHN [33]	5.6	-	-	-
DCFE	5.27	35.86	7.29	81.59/49.57

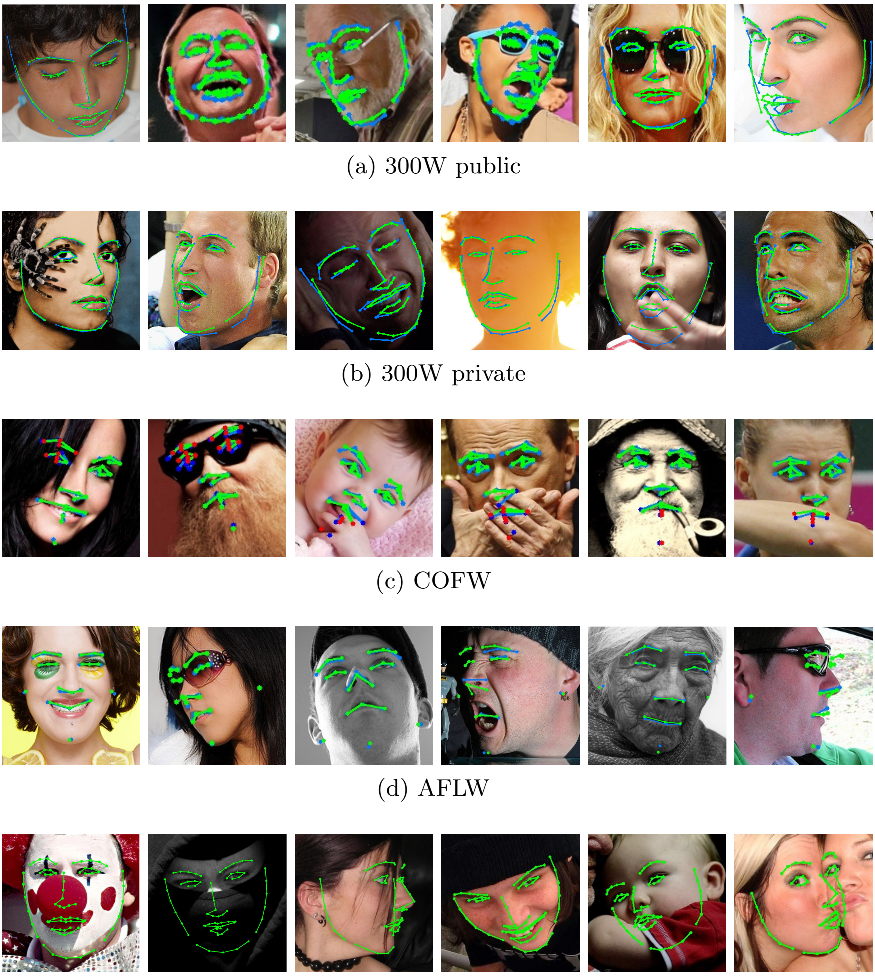
Table 3: COFW results.

Method	height NME
ESR [7]	4.35
CFSS [38]	3.92
RCPR [6]	3.73
Bulat *et al.* [5]	2.85
CCL [37]	2.72
DAC-CSR [14]	2.27
TSR [23]	2.17
DCFE	2.17

Table 4: AFLW results.

Menpo test annotations have not been released, but we have processed their testing images to visually perform an analysis of the errors. In comparison with many other approaches our algorithm evaluates in both subsets training a unique semi-supervised model through the 68 (semi-frontal) and 39 (profile) landmark annotations all together. We detect test faces using the public Single Shot Detector [22] from OpenCV. We manually filter the detected face bounding boxes to reduce false positives and improve the accuracy.

In Fig. 5 we present some qualitative results for all data sets, including Menpo.



(a) 300W public

(b) 300W private

(c) COFW

(d) AFLW

(e) Menpo

Fig. 5: Representative results using DCFE in 300W, COFW, AFLW and Menpo testing subsets. Blue colour represents ground truth, green and red colours point out visible and non-visible shape predictions respectively.

5 Conclusions

In this paper we have introduced DCFE, a robust face alignment method that leverages on the best features of the three main approaches in the literature: 3D face models, CNNs and ERT. The CNN provides robust landmark estimations with no face shape enforcement. The ERT is able to enforce face shape and achieve better accuracy in landmark detection, but it only converges when properly initialized. Finally, 3D models exploit face orientation information to improve self-occlusion estimation. DCFE combines CNNs and ERT by fitting a 3D model to the initial CNN prediction and using it as initial shape of the ERT. Moreover, the 3D reasoning capability allows DCFE to easily handle self occlusions and deal with both frontal and profile faces.

Once we have solved the problem of ERT initialization, we can exploit its benefits. Namely, we are able to train it in a semi-supervised way with missing landmarks. We can also estimate landmark visibility due to occlusions and we can parallelize the execution of the regression trees in each stage. We have additionally introduced a coarse-to-fine ERT that is able to deal with the combinatorial explosion of local parts deformation. In this case, the usual monolithic ERT will perform poorly when fitting faces with combinations of facial part deformations not present in the training set.

In the experiments we have shown that DCFE runs in real-time improving, as far as we know, the state-of-the-art performance in 300W, COFW and AFLW data sets. Our approach is able to deal with missing and occluded landmarks allowing us to train a single regressor for both full profile and semi-frontal images in the Mempo and AFLW data sets.

Acknowledgments: The authors thank Pedro López Maroto for his help implementing the CNN. They also gratefully acknowledge computing resources provided by the Super-computing and Visualization Center of Madrid (CeSViMa) and funding from the Spanish Ministry of Economy and Competitiveness under project TIN2016-75982-C2-2-R. José M. Buenaposada acknowledges the support of Computer Vision and Image Processing research group (CVIP) from Universidad Rey Juan Carlos.

References

1. Alahi, A., Ortiz, R., Vandergheynst, P.: FREAK: fast retina keypoint. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
2. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Robust gender recognition by exploiting facial attributes dependencies. Pattern Recognition Letters (PRL) **36**, 228–234 (2014)
3. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
4. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) (2003)
5. 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)
6. Burgos-Artizzu, X.P., Perona, P., Dollar, P.: Robust face landmark estimation under occlusion. In: Proc. International Conference on Computer Vision (ICCV) (2013)
7. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Proc. European Conference on Computer Vision (ECCV) (1998)
9. Dantone, M., Gall, J., Fanelli, G., Gool, L.V.: Real-time facial feature detection using conditional regression forests. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
10. David, P., DeMenthon, D., Duraiswami, R., Samet, H.: Softposit: Simultaneous pose and correspondence determination. International Journal of Computer Vision (IJCV) **59**(3), 259–284 (2004)
11. Deng, J., Liu, Q., Yang, J., Tao, D.: CSR: multi-view, multi-scale and multi-component cascade shape regression. Image and Vision Computing (IVC) **47**, 19–26 (2016)
12. Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
13. Fan, H., Zhou, E.: Approaching human level facial landmark localization by deep learning. Image and Vision Computing (IVC) **47**, 27–35 (2016)
14. Feng, Z., Kittler, J., Christmas, W.J., Huber, P., Wu, X.: Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
15. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer (2009)
16. 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)
17. Jourabloo, A., Ye, M., Liu, X., Ren, L.: Pose-invariant face alignment with a single CNN. In: Proc. International Conference on Computer Vision (ICCV) (2017)
18. 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)

19. Kowalski, M., Naruniec, J.: Face alignment using k-cluster regression forests with weighted splitting. IEEE Signal Processing Letters **23**(11), 1567–1571 (2016)
20. 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)
21. 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)
22. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: Proc. European Conference on Computer Vision (ECCV) (2016)
23. 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)
24. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
25. Soltanpour, S., Boufama, B., Wu, Q.M.J.: A survey of local feature methods for 3D face recognition. Pattern Recognition (PR) **72**, 391–406 (2017)
26. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
27. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
28. Wu, Y., Ji, Q.: Robust facial landmark detection under significant head poses and occlusion. In: Proc. International Conference on Computer Vision (ICCV) (2015)
29. Xiao, S., Feng, J., Liu, L., Nie, X., Wang, W., Yan, S., Kassim, A.A.: Recurrent 3d-2d dual learning for large-pose facial landmark detection. In: Proc. International Conference on Computer Vision (ICCV) (2017)
30. 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)
31. Xiong, X., la Torre, F.D.: Supervised descent method and its applications to face alignment. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
32. Xiong, X., la Torre, F.D.: Global supervised descent method. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
33. 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)
34. Yu, X., Zhou, F., Chandraker, M.: Deep deformation network for object landmark localization. In: Proc. European Conference on Computer Vision (ECCV) (2016)
35. 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)
36. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Proc. European Conference on Computer Vision (ECCV) (2014)

37. 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)
38. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)