Face Alignment using a 3D Deeply-initialized Ensemble of Regression Trees

Roberto Valle[a], José M. Buenaposada[b,**], Antonio Valdés[c], 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
[c]Departamento de Geometría y Topología in Universidad Complutense, Plaza Ciencias, 3, 28040 Madrid, Spain

ABSTRACT

Face alignment algorithms locate a set of landmark points in images of faces taken in unrestricted situations. State-of-the-art approaches typically fail or lose accuracy in the presence of occlusions, strong deformations, large pose variations and ambiguous configurations. In this paper we present 3DDE, a robust and efficient face alignment algorithm based on a coarse-to-fine cascade of ensembles of regression trees. It is initialized by robustly fitting a 3D face model to the probability maps produced by a convolutional neural network. With this initialization we address self-occlusions and large face rotations. Further, the regressor implicitly imposes a prior face shape on the solution, addressing occlusions and ambiguous face configurations. Its coarse-to-fine structure tackles the combinatorial explosion of parts deformation. In the experiments performed, 3DDE improves the state-of-the-art in 300W, COFW, AFLW and WFLW data sets. Finally, we perform cross-dataset experiments that reveal the existence of a significant data set bias in these benchmarks.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Face alignment algorithms precisely locate a set of points of interest in the images of faces taken in unrestricted conditions. It has received much attention from the research community (Jin and Tan, 2017) since it is a preliminary step for estimating 3D facial structure (Zhao et al., 2016) and many other face image analysis problems such as verification and recognition (Soltanpour et al., 2017), attributes estimation (Bekios-Calfa et al., 2014) or facial expression recognition (Martinez and Du, 2012), to name a few. Present approaches typically fail or lose precision in the presence of occlusions, strong deformations produced by facial expressions, large pose variations and ambiguous configurations caused, for example, by strong make-up or the existence of other nearby faces.

Top performers in the most popular benchmarks are based on Convolutional Neural Networks (CNNs) and Ensemble of Regression Trees (ERT), see *e.g*., Tables 1, 2, 3, 4 and 5. The large effective receptive field of deep models (Kowalski et al., 2017; Lv et al., 2017; Xiao et al., 2016; Yang et al., 2017; Wu et al., 2018) enable them to model context better and produce robust landmark estimations. However, in these models it is not easy to enforce facial shape consistency, something that limits their accuracy in the presence of occlusions and ambiguous facial configurations. ERT-based models (Burgos-Artizzu et al., 2013; Cao et al., 2014; Kazemi and Sullivan, 2014; Lee et al., 2015; Ren et al., 2016), on the other hand, are difficult to initialize, but may implicitly impose face shape consistency in their estimations (Cao et al., 2014). This increases their performance in occluded and ambiguous situations. They are also

---

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

much more efficient than deep models and, as we demonstrate in our experiments, with a good initialization they are also very accurate.

In this paper we present the 3DDE (3D Deeply-initialized Ensemble) regressor, a robust and efficient face alignment algorithm based on a coarse-to-fine cascade of ERTs. It is a hybrid approach that inherits good properties of ERT, such as the ability to impose a face shape prior, and the robustness of deep models. It is initialized by robustly fitting a 3D face model to the probability maps produced by a CNN. With this initialization we tackle one of the main drawbacks of ERT, namely the difficulty in initializing the regressor in the presence of occlusions and large face rotations. On the other hand, the ERT implicitly imposes a prior face shape on the solution, addressing the shortcomings of deep models when occlusions and ambiguous face configurations are present. Finally, its coarse-to-fine structure tackles the combinatorial explosion of parts deformation, which is also a key limitation of approaches using shape constraints (Cao et al., 2014).

A preliminary version of our work appeared in Valle et al. (2018). Here we refine and extend it in several ways. First we improve the initialization by using a RANSAC-like procedure that increases its robustness in the presence of occlusions. We have also introduced early stopping and better data augmentation techniques for increasing the regularization when training both the ERT and the CNN. We also extend the evaluation including the newly released WFLW data base and a detailed ablation study. Finally, 3DDE may also be trained in presence of missing and occluded landmarks in the training set. This has enabled us to perform cross-dataset experiments that reveal the existence of significant data set bias that may limit the generalization capabilities of regressors trained on present data bases. To the best of our knowledge, this is the first time such a problem has been raised in the field.

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) (Cootes et al., 1998) or the 3D Morphable Models (3DMM) (Blanz and Vetter, 2003). Recent approaches are based on a cascaded combination of discriminative regressors.

In the earliest case these regressors are Random Ferns (Dollar et al., 2010), Ensembles of Regression Trees (Cao et al., 2012) or linear models (Xiong and la Torre, 2013, 2015). Key ideas in this approach are indexing image description relative to the current shape estimate (Dollar et al., 2010), and the use of a regressor whose predictions lie on the subspace spanned by the training face shapes (Cao et al., 2014), this is the so-called Cascade Shape Regressor (CSR) framework. Kazemi and Sullivan (2014) improved the original cascade framework by proposing a real-time ensemble of regression trees. Ren et al. (2016) used locally binary features to boost the performance up to 3000 FPS. Burgos-Artizzu et al. (2013) included occlusion estimation and decreased the influence of occluded landmarks. Shen et al. (2014) refine the initial location of face landmarks using a random forest and SIFT features. Xiong and la Torre (2013, 2015) also use SIFT features and learn the linear regressor dividing the search space into individual regions with similar gradient directions. Overall, this set of approaches are very sensitive to the starting point of the regression process. For this reason an important part of recent work revolves around how to find good initializations (Zhu et al., 2015, 2016). However, they are extremely efficient and may take advantage of implicit shape constraints (Cao et al., 2012, 2014).

The recent development of deep learning techniques has also impacted the face alignment field with the widespread use of CNN-based regressors. Sun et al. (2013) were pioneers to apply a three-level CNN for locating landmarks. Zhang et al. (2014) proposed a multi-task solution to deal with face alignment and attributes classification. Lv et al. (2017) use global and local face parts regressors for fine-grained facial deformation estimation. Yu et al. (2016) transform the landmarks rather than the input image for the refinement cascade. Trigeorgis et al. (2016) and Xiao et al. (2016) are the first approaches that fuse the feature extraction and regression steps of CSR into a recurrent neu-

ral network trained end-to-end. 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. Wu et al. (2018) derive face landmarks from boundary lines, which helps to remove the ambiguities in the landmark definition. Deep CNN models have large effective receptive fields that let them model context better and convey these approaches with a high degree of robustness to face rotation, scale, deformation and initialization. However, when used in a cascaded framework they may notably increase the computational requirements. Moreover, it is not clear how to impose facial shape consistency on the estimated set of landmarks. Hence, the regressor accuracy may be harmed in the presence of occlusions or ambiguities.

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 positions (Kowalski and Naruniec, 2016) or jointly regress the pose and shape of the face (Jourabloo et al., 2017; Xiao et al., 2017). Zhu et al. (2017) and Kumar and Chellappa (2018) fit a 3DMM in a cascaded way. 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 on the good properties of the three approaches described above. Using a CNN-based initialization we inherit the robustness of deep learning models. Like the simple 3D approaches we fit a rigid 3D face model to initialize the regressor and estimate the initial face orientation to address self-occlusions and ambiguities. Finally, we use a cascaded ERT within a coarse-to-fine framework to achieve accuracy and efficiency while avoiding the combinatorial explosion of independent parts deformations.

3. 3D deeply-initialized Ensemble

In this section we present 3DDE. 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.

3.1. Rigid pose computation

ERT-based regressors require a good initialization to converge. We propose the use of face landmarks location probability maps (Belhumeur et al., 2011; Dantone et al., 2012; Xiao et al., 2016) to generate plausible shape initialization candidates. We define a UNet-like architecture (Ronneberger et al., 2015; Honari et al., 2016), with a loss function that handles missing landmarks. We train this CNN to obtain a set of probability maps, $\mathcal{P}(\mathbf{I})$, that model 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, ambiguities and may not be a valid face shape. Compared to typical CNN-based approaches, *e.g.*, Yang et al. (2017), our CNN is much 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. (2004) within a robust scheme. Unlike Valle et al. (2018), here we use a set of the distinct landmarks to establish the correspondences between the CNN predictions and the 3D face model. This avoids problems related to ambiguous landmarks around the jaw that do not correspond always to the same 3D points and produce wrong initializations, mainly in profile faces. Moreover, we have also implemented a RANSAC-like procedure, that runs softPOSIT several times with subsets of correspondences, to obtain a robust estimation (see Algorithm 1).

Let $\mathbf{X} \in \mathbb{R}^{L \times 3}$ be the 3D coordinates of the L landmarks on the 3D face model, $\mathbf{x} \in \mathbb{R}^{L \times 2}$ their 2D projections onto the image plane and $\mathbf{v} \in \{0, 1\}^L$ their visibilities. We produce subsets of correspondences $(\mathbf{x}_s, \mathbf{X}_s)$ from the distinct landmarks shown
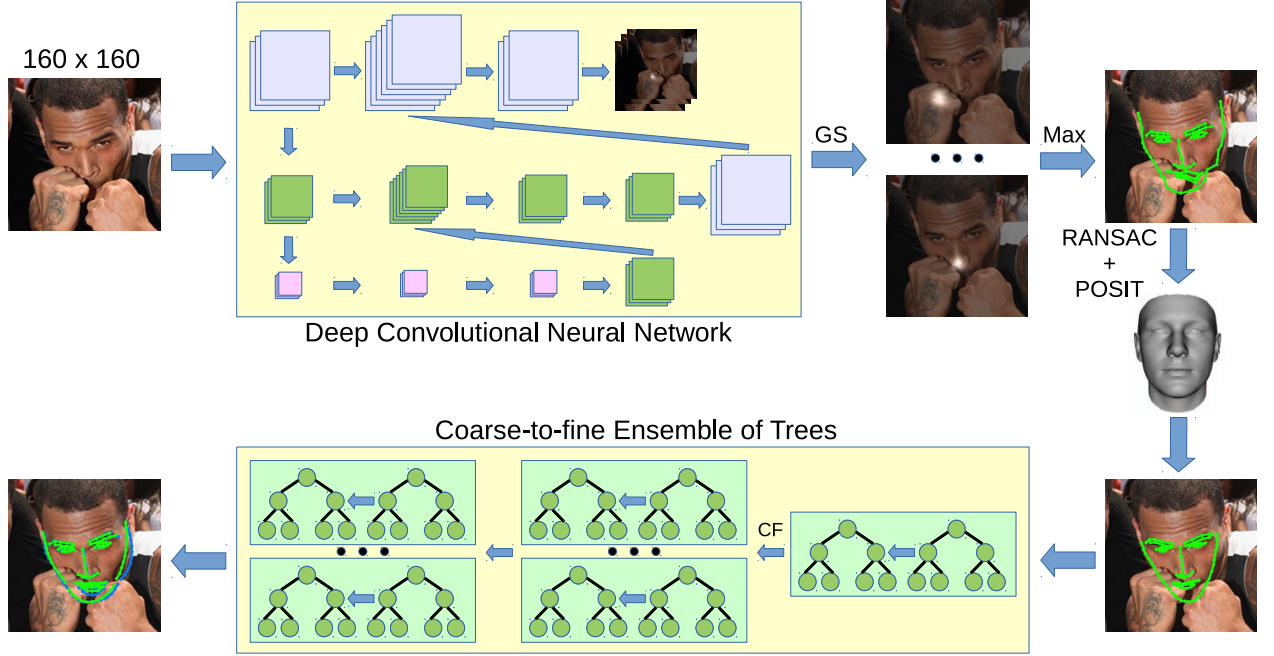
Fig. 1. 3DDE framework diagram. GS, Max and RANSAC+POSIT represent the Gaussian smoothing filter, the maximum of each probability map and the robust 3D pose estimation respectively.

Algorithm 1 Initialization algorithm (g_0)

Input: $\mathcal{P}(\mathbf{I})$, $\mathbf{X}$

    // Select coordinates of maximum probability

    $\{\mathbf{x}(l) = \arg\max(\mathcal{P}^l(\mathbf{I}))\}_{l=1}^L$

    $p^* = 0$

    for z=1 **to** Z **do**

        // Select subset from distinct landmarks

        $\mathbf{x}_s, \mathbf{X}_s = \text{chooseLandmarksSubset}(\mathbf{x}, \mathbf{X})$

        // Compute projection matrix between $\mathbf{x}_s, \mathbf{X}_s$

        $\mathbf{R}, \mathbf{t} = \text{softPOSIT}(\mathbf{x}_s, \mathbf{X}_s)$

        // Project 3D face model using previous matrix

        $\mathbf{x}_z, \mathbf{v}_z = \text{projectPoints}(\mathbf{X}, \mathbf{R}, \mathbf{t})$

        // Evaluate the goodness of the initialization

        $p(\mathbf{x}_z) = \sum_{l=1}^L \mathcal{P}^l(\mathbf{I})[\mathbf{x}_z(l)]$

        if $p(\mathbf{x}_z) > p^*$ **then**

            $p^* = p(\mathbf{x}_z), \mathbf{R}^* = \mathbf{R}, \mathbf{t}^* = \mathbf{t}$

        end if

    end for

    $\mathbf{x}^0, \mathbf{v}^0 = \text{projectPoints}(\mathbf{X}, \mathbf{R}^*, \mathbf{t}^*)$

Output: $\mathbf{x}^0, \mathbf{v}^0$

---

in Fig. 8a, estimate the 3D face model pose $(\mathbf{R}, \mathbf{t})$ with soft-POSIT and evaluate the goodness of each estimation as the sum of landmarks probabilities,

$$p(\mathbf{x}_z) = \sum_{l=1}^L \mathcal{P}^l(\mathbf{I})[\mathbf{x}_z(l)],$$

where $\mathbf{x}_z(l)$ are the 2D coordinates of the l-th landmark and $\mathcal{P}^l(\mathbf{I})$ is the probability map for landmark l. Finally, we select the rigid transformation $(\mathbf{R}, \mathbf{t})$ with highest $p(\mathbf{x}_z)$. As a result, we project the 3D model onto the image using the most likely 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), and the visibility estimation of the self-occluded parts of the face.

Let $\mathbf{x}^0 = g_0(\mathcal{P}(\mathbf{I}), \mathbf{X})$ be the *initial shape*, the output of the initialization function g_0 after processing the input image $\mathbf{I}$. With our initialization we enforce two key requirements for the convergence of the ERT. First, that $\mathbf{x}^0$ lies on the face with an approximately correct 3D face pose. Second, that $\mathbf{x}^0$ is a valid face shape. The latter guarantees that the predictions in the next step of the algorithm will also be valid face shapes (Cao et al., 2014).

3.2. ERT-based non-rigid shape estimation

Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ be the set of training face shapes, where $s_i = (\mathbf{I}_i, \mathbf{x}_i^g, \mathbf{v}_i^g, \mathbf{w}_i^g, \mathbf{x}_i^0, \mathbf{v}_i^0)$. Each shape s_i has its own training image, $\mathbf{I}_i$, ground truth shape, $\mathbf{x}_i^g$, ground truth visibility label, $\mathbf{v}_i^g$, annotated landmark label, $\mathbf{w}_i^g \in \{0,1\}^L$, initial shape, $\mathbf{x}_i^0$, and visibilities, $\mathbf{v}_i^0$, for training the ERT regressor. In our implementation we use shape-indexed features (Lee et al., 2015), $\phi(\mathcal{P}(\mathbf{I}_i), \mathbf{x}_i^t, \mathbf{w}_i^g)$, that depend on the current shape $\mathbf{x}_i^t$ of the landmarks in image $\mathbf{I}_i$ and whether they are annotated or not, $\mathbf{w}_i^g$.

We divide the regression process into a maximum of T stages. We learn an ensemble of K regression trees for the t-th stage, $C_t(f_i) = \mathbf{x}_i^{t-1} + \sum_{k=1}^{K} g_k(f_i)$, where $f_i = \phi(\mathcal{P}(\mathbf{I}_i), \mathbf{x}_i^{t-1}, \mathbf{w}_i^g)$ and $\mathbf{x}^j$ are the coordinates of the landmarks estimated in j-th stage. To train the ERT we use the N training shapes in $\mathcal{S}$ to generate an augmented training set of samples, $\mathcal{S}_A$, and a validation set, $\mathcal{S}_V$, with cardinality $N_A = |\mathcal{S}_A|$ and $N_V = |\mathcal{S}_V|$ respectively. The total number of samples is $N_T = N_A + N_V$. Instead of using a fixed number of stages, like Valle et al. (2018), we stop training when the validation error stops improving. In this way the regressor has a variable number of stages. We compute the initialization for each sample using the 3D projections produced by g_0 (see generated initializations in Fig. 2). We also improve the data augmentation used in Valle et al. (2018). To this end we add random noise to the yaw, pitch and roll angles, of the rotation matrix $\mathbf{R}^*$ estimated with g_0, to generate new training initializations for each sample in $\mathcal{S}_A$.

Following *et al.* Burgos-Artizzu et al. (2013) and Kazemi and Sullivan (2014), we attach to each landmark in $\mathcal{S}$ the binary labels $\{\mathbf{v}, \mathbf{w}\} \in \{0, 1\}$ that model respectively whether it is visible and annotated. We learn these labels in the ERT together with the landmark location. Each initial shape is progressively refined by estimating a shape and visibility increments $C_t^{\mathbf{v}}(\phi(\mathcal{P}(\mathbf{I}_i), \mathbf{x}_i^{t-1}, \mathbf{w}_i^g))$ where $\mathbf{x}_i^{t-1}$ represents the current shape of the i-th sample (see Algorithm 2). $C_t^{\mathbf{v}}$ 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 $\mathcal{A}_{t-1} = \{(\mathbf{x}_i^{t-1}, \mathbf{v}_i^{t-1})\}_{i=1}^{N_A}$ and $\mathcal{V}_{t-1} = \{(\mathbf{x}_i^{t-1}, \mathbf{v}_i^{t-1})\}_{i=1}^{N_V}$ as the set of all current shapes and corresponding visibility vectors for all training and validation data, respectively.

Algorithm 2 Training an Ensemble of Regression Trees

Input: $\mathcal{S}, T$

   // Generate an augmented training set of samples

   $\mathcal{S}_A, \mathcal{S}_V = \text{dataAugmentation}(\mathcal{S})$

repeat

   // Extract training ($\mathcal{F}_A$) and validation ($\mathcal{F}_V$) features

   $\mathcal{F}_A \cup \mathcal{F}_V = \{f_i\}_{i=1}^{N_T} = \{\phi(\mathcal{P}(\mathbf{I}_i), \mathbf{x}_i^{t-1}, \mathbf{w}_i^g)\}_{i=1}^{N_T}$

   // Apply Algorithm 3 using training samples

   $C_t^{\mathbf{v}} = \text{learnCoarseToFineRegressor}(\mathcal{S}_A, \mathcal{F}_A, \mathcal{A}_{t-1}, K, P)$

   // Update validation samples

   $\mathcal{V}_t = \mathcal{V}_{t-1} + \{C_t^{\mathbf{v}}(f_i)\}_{i=1}^{N_V}$

   // Increase P when $NME(\{\mathbf{x}_i^t, \mathbf{x}_i^g\}_{i=1}^{N_A}) < NME(\{\mathbf{x}_i^t, \mathbf{x}_i^g\}_{i=1}^{N_V})$

   // Compute validation error improvement

   $\Delta\varepsilon = NME(\{\mathbf{x}_i^{t-1}, \mathbf{x}_i^g\}_{i=1}^{N_V}) - NME(\{\mathbf{x}_i^t, \mathbf{x}_i^g\}_{i=1}^{N_V})$

until t > T **or** $\Delta\varepsilon < 1\%$

Output: $\{C_t^{\mathbf{v}}\}_{t=1}^{T^*}$ // T^* is the last trained stage

---

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

3.2.1. 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, $\bar{\mathbf{x}}^0 = \sum_{i=1}^{N} \mathbf{x}_i^g / N$. However, such a poor initialization leads to wrong alignment results in test images with large pose variations. Alternative strategies run the ERT several times with different initializations (Burgos-Artizzu et al., 2013), initialize with other ground truth shapes $\mathbf{x}_i^0 \leftarrow \mathbf{x}_j^g$ where $i \neq j$ (Kazemi and Sullivan, 2014), or randomly deform the initial shape (Kowalski et al., 2017).

In our approach we initialize the ERT using the algorithm described in section 3.1, that provides a robust pose and a valid
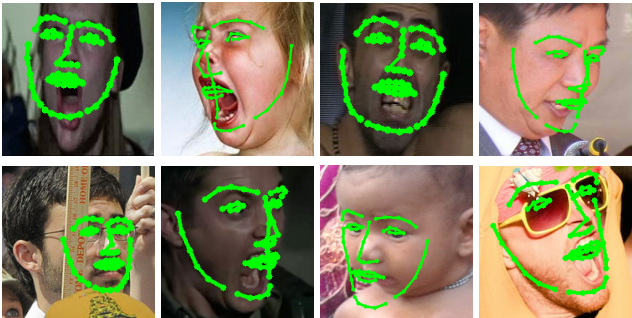
Fig. 2. The 8 worst initial shapes for the 300W private test set produced by g_0 (CNN+3D).

Fig. 3. The $P = 10$ face parts of 300W, COFW, AFLW and WFLW data bases in the fine stage of our coarse-to-fine ERT.

shape for initialization (see Fig. 2). Hence, the ERT only needs to estimate the non-rigid deformation component of the face.

3.2.2. Feature Extraction

ERT efficiency depends on the feature extraction step. In general, descriptor features such as SIFT used by Xiong and la Torre (2013) and Zhu et al. (2015) improve face alignment results, but have higher computational cost compared to simpler features such as plain pixel value differences (Cao et al., 2014; Burgos-Artizzu et al., 2013; Kazemi and Sullivan, 2014; Ren et al., 2016). In our case, a simple feature suffices, since shape landmarks are close to their ground truth location.

We use the probability maps $\mathcal{P}(\mathbf{I})$ to extract features for the cascade. To this end, we select a landmark l and its associated probability map $\mathcal{P}^l(\mathbf{I})$. The feature is computed as the difference between two pixels values in $\mathcal{P}^l(\mathbf{I})$ from a FREAK descriptor pattern (Alahi et al., 2012) around l, similar to those in Lee et al. (2015). However, ours are defined on the probability maps, $\mathcal{P}(\mathbf{I})$, instead of the image, $\mathbf{I}$. We let the training algorithm select the most informative landmark and pair of pixels in each iteration.

3.2.3. Learn a coarse-to-fine regressor

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

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

where $\odot$ is the Hadamard product. There are different ways of minimizing Equation 1. Kazemi and Sullivan (2014) present a
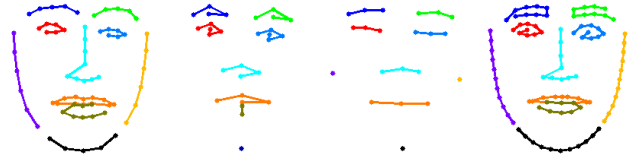
general framework based on Gradient Boosting for learning an ensemble of regression trees. Lee et al. (2015) establish an optimization method based on Gaussian Processes also learning an ensemble of regression trees but outperforming previous literature by reducing the overfitting. In our approach we adopt a Gradient Boosting scheme (see Algorithm 3).

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 the difference between monolithic and coarse-to-fine NME curves in Fig. 6a). Our algorithm is agnostic in the number of parts and stages of the coarse-to-fine estimation. Algorithm 3 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 $\mathcal{A}_{k-1}$ in this context is 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 the coarse-to-fine scheme has two stages. The coarse stage has one part, $P = 1$, that involves all landmarks and K_1 trees. The fine stage has ten parts, $P = 10$, left/right eyebrow, left/right eye, nose, top/bottom mouth, left/right ear and chin (see Fig. 3), with K_2 trees.

3.2.4. 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 an-

---

Algorithm 3 Training P parts regressors

Input: $\mathcal{S}_A, \mathcal{F}_A, \mathcal{A}_{t-1}, v, K, P$

    for k=1 **to** K **do**

        for p=1 **to** P **do**

            // Compute residuals:

            // $\odot$ is the Hadamard product

            // (p) selects elements of vectors in that part

            $\{\mathbf{r}_i^k(p) = \mathbf{w}_i^g(p) \odot (\mathbf{x}_i^g(p) - \mathbf{x}_i^{k-1}(p))\}_{i=1}^{N_A}$

            $g_k^p = \text{fitRegressionTree}(\{\mathbf{r}_i^k(p)\}_{i=1}^{N_A}, \mathcal{F}_A(p))$

            // Update samples with the regression tree estimation,

            // v, shrinkage factor to scale each tree contribution

            $\mathcal{A}_k(p) = \mathcal{A}_{k-1}(p) + v \cdot \{g_k^p(f_i(p))\}_{i=1}^{N_A}$

        end for

    end for

Output: $\{C^p\}_{p=1}^P$, being $C^p = \{g_k^p\}_{k=1}^K$

---

notated 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 $f_i(\theta) = \mathcal{P}^l(\mathbf{I}_i)[\mathbf{p}_1] - \mathcal{P}^l(\mathbf{I}_i)[\mathbf{p}_2]$, 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. We also output the mean visibility of the samples reaching the tree leaf.

4. Experiments

To train and evaluate our proposal, we perform experiments with 300W, COFW, AFLW and WFLW that are considered the most challenging public data sets:

- **300W**. It provides 68 manually annotated landmarks, Sagonas et al. (2016). 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 300W private competition using the previous 3837 images as training and 600 newly updated images as testing set.

- **COFW**. This benchmark, presented in Burgos-Artizzu et al. (2013) 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**. It provides a collection of 25993 in-the-wild faces, with 21 facial landmarks annotated depending on their visibility, Koestinger et al. (2011). 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.

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

4.1. Evaluation

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

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

It computes the mean 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 ground truth distance between the eye centers (*pupils*), the ground truth distance between the outer eye corners (*corners*) and the ground truth 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 of each data base 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. For each data set, we train from scratch the CNN selecting the model parameters with lowest validation error. We crop faces using the ground truth bounding boxes annotations enlarged by 30%. We generate different training samples in each epoch by applying random in plane rotations between ±45°, 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 until convergence with an initial learning rate $\alpha = 0.001$. When validation error levels out for 10 epochs, we multiply the learning rate by *decay* = 0.05. 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 stride 2 convolution with kernel size 2×2[1]. We apply

---

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

batch normalization after each convolution. All layers contain 68 filters to describe the required landmark features. We apply a Gaussian filter with $\sigma = 33$ to the output probability maps to stabilize the initialization, g_0.

We train the coarse-to-fine ERT with the Gradient Boosting algorithm (Hastie et al., 2009). It requires a maximum of $T = 20$ stages of $K = 50$ regression trees per stage. The depth of trees is set to 4. 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 160×160 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 $Z = 25$ initializations in the robust softPOSIT scheme of g_0. We augment the shapes of each face training image to create a set, $\mathcal{S}_A$, of at least $N_A = 60000$ samples to train the cascade. To avoid overfitting we use a shrinkage factor $\nu = 0.1$ and subsampling factor $\eta = 0.5$ in the ERT. Our regressor triggers the coarse-to-fine strategy once the training error is below the validation error, *e.g.*, $t = 5$ in Fig. 6a.

Training the CNN and the coarse-to-fine ensemble of trees takes 48 hours using a NVidia GeForce GTX 1080Ti (11GB) GPU and an dual Intel Xeon Silver 4114 CPU at 2.20GHz (2×10 cores/20 threads, 128 GB of RAM) with a batch size of 32 images. At runtime our method process test images on average at a rate of 12.5 FPS, where the CNN takes 75 ms and the ERT 5 ms per face image using C++, Tensorflow and OpenCV libraries.

4.3. Experiments using public code

Published results in the literature are sometimes not fully comparable. In this section we use publicly available code to ensure a fair comparison between 3DDE and DCFE (Valle et al., 2018), LAB (Wu et al., 2018), DAN (Kowalski et al., 2017), RCN (Honari et al., 2016), cGPRT (Lee et al., 2015), RCPR (Burgos-Artizzu et al., 2013) and ERT (Kazemi and Sullivan, 2014) with the same settings (including same training, validation and bounding boxes), in different benchmarks: 300W public, 300W private, COFW and WFLW. Note that LAB (Wu et al., 2018) only provides a trained model for the
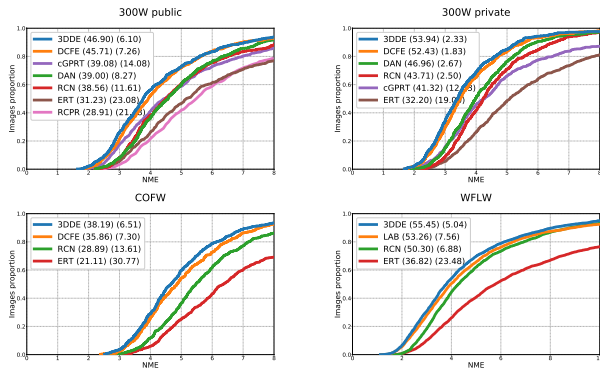
Fig. 4. Cumulative error distribution curves sorted by AUC.

WFLW data set. In addition, DAN (Kowalski et al., 2017) provides code using 68 landmarks, for this reason we only report results in 300W. In Fig. 4 we plot the CED curves for all data bases. In the legend we provide the *AUC* and *FR* values for each algorithm.

The selected algorithms are representative of the three main families of solutions: a) ensembles of regression trees (cGPRT, RCPR, ERT), b) CNN-based approaches (LAB, DAN, RCN) and c) mixed approaches with deep nets and ensembles of regression trees (3DDE, DCFE). Overall, 3DDE is better than any other providing a public implementation in the literature. We improve over our preliminary algorithm, DCFE Valle et al. (2018), because of the better 3D initialization and regularization (see a complete analysis in section 4.5). In general we are able to improve by a large margin other ERT methods as RCPR, ERT or cGPRT because of the better initialization and the robust features provided by the CNN. We also outperform RCN (without any denoising model), a CNN architecture like the one used in 3DDE. Even DAN and LAB, that implement a cascade of CNN regressors, can not compete with the regularization obtained by using the cascade of ERT in 3DDE (see Fig. 4). The fact that the largest margin is in COFW reflects the importance of the implicit shape model in our cascade to address occlusions.

4.4. Experiments using published results

In this section we compare 3DDE with other methods in the literature by using their published results. Since our method is

Method	Common		Challenging		Full			
	pupils	corners	pupils	corners	pupils	corners		
	NME	NME	NME	NME	NME	NME	AUC_8	FR_8
RCPR (Burgos-Artizzu et al., 2013)	6.18	-	17.26	-	8.35	-	-	-
ESR (Cao et al., 2012)	5.28	-	17.00	-	7.58	-	43.12	10.45
SDM (Xiong and la Torre, 2013)	5.60	-	15.40	-	7.52	-	42.94	10.89
ECSAN (Zhang and Hu, 2018)	5.42	-	11.80	-	6.67	-	-	-
ERT (Kazemi and Sullivan, 2014)	-	-	-	-	6.40	-	-	-
LBF (Ren et al., 2016)	4.95	-	11.98	-	6.32	-	-	-
cGPRT (Lee et al., 2015)	-	-	-	-	5.71	-	-	-
CFSS (Zhu et al., 2015)	4.73	-	9.98	-	5.76	-	49.87	5.08
DDN (Yu et al., 2016)	-	-	-	-	5.65	-	-	-
TCDCN (Zhang et al., 2014)	4.80	-	8.60	-	5.54	-	-	-
MDM (Trigeorgis et al., 2016)	-	-	-	-	-	-	52.12	4.21
3DDFA (Zhu et al., 2017)	5.09	-	8.07	-	5.63	-	-	-
RCN (Honari et al., 2016)	4.67	-	8.44	-	5.41	-	-	-
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	-	-	-
SHN (Yang et al., 2017)	4.12	-	**7.00**	**4.90**	4.68	-	-	-
DCFE (Valle et al., 2018)	3.83	2.76	7.54	5.22	4.55	3.24	60.13	1.59
PCD-CNN (Kumar and Chellappa, 2018)	**3.67**	-	7.62	-	4.44	-	-	-
3DDE	3.73	**2.69**	7.10	4.92	**4.39**	**3.13**	**61.24**	1.30

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

able to train with unannotated landmarks and visibilities, we are able to train and evaluate all data sets in the literature.

First we test our method against the 300W benchmark. Our approach obtains the best overall performance in the indoor and outdoor subsets of the private competition (see Table 2) and in the full subset of the 300W 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 and the deep robust features extracted from the CNN. In the challenging subset of the 300W public competition, SHN (Yang et al., 2017) gets better results than 3DDE. This is due to 3DDE failing to estimate good landmark probability maps for images with large scale variations. Our method exhibits superior capability in handling typical cases in the data base, since we achieve the best NME full set results in 300W public, 4.39, and in 300W private, 3.73.

We may assess the improvement achieved by the 3D initialization and the coarse-to-fine ERT by comparing the results of 3DDE in the full subset of 300W, 4.39, with Honari's RCN using the denoising model (Honari et al., 2016), 5.41. It roughly represents a 19% improvement in the inter-pupils NME.

Method	Indoor corners			Outdoor corners			Full corners		
	NME	AUC_8	FR_8	NME	AUC_8	FR_8	NME	AUC_8	FR_8
ESR (Cao et al., 2012)	-	-	-	-	-	-	-	32.35	17.00
cGPRT (Lee et al., 2015)	-	-	-	-	-	-	-	41.32	12.83
CFSS (Zhu et al., 2015)	-	-	-	-	-	-	-	39.81	12.30
MDM (Trigeorgis et al., 2016)	-	-	-	-	-	-	5.05	45.32	6.80
DAN (Kowalski et al., 2017)	-	-	-	-	-	-	4.30	47.00	2.67
SHN (Yang et al., 2017)	4.10	-	-	4.00	-	-	4.05	-	-
DCFE (Valle et al., 2018)	3.96	52.28	2.33	3.81	52.56	**1.33**	3.88	52.42	**1.83**
3DDE	**3.74**	**53.93**	**2.00**	**3.71**	**53.95**	2.66	**3.73**	**53.94**	2.33

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

Method	pupils			occlusion
	NME	AUC_8	FR_8	precision/recall
RCPR (Burgos-Artizzu et al., 2013)	8.50	-	-	80/40
TCDCN (Zhang et al., 2014)	8.05	-	-	-
RAR (Xiao et al., 2016)	6.03	-	-	-
DAC-CSR (Feng et al., 2017)	6.03	-	-	-
Wu *et al.* (Wu and Ji, 2015)	5.93	-	-	80/49.11
SHN (Yang et al., 2017)	5.6	-	-	-
PCD-CNN (Kumar and Chellappa, 2018)	5.77	-	-	-
DCFE (Valle et al., 2018)	5.27	35.86	7.29	81.59/49.57
3DDE	**5.11**	**38.18**	**6.50**	**85.92/51.04**

Table 3. Error of face alignment methods on COFW.

Table 3 compares the performance of our model using the COFW data set. This is the standard to evaluate occlusions. 3DDE obtains the best results, NME 5.11, establishing a new state-of-the-art. This shows the importance of the face shape model implicit in the cascade of ERT to cope with severe occlusions. In terms of landmark visibility estimation, we have obtained better precision with an overall better recall than the best previous approach, DCFE. Again, the regularization together with the new initialization contributes to improve DCFE.

In Table 4 we show the results of our evaluation with AFLW. This is a challenging data set not only because of its size and the large variability of face poses, but also because of the large number of samples with occluded landmarks, that are unannotated. Although the results in Table 4 are not strictly comparable, because each paper uses its own train and test subsets, we get an NME of 2.06 with the full 21 landmarks set. Again, it is a new state-of-the-art, since most competing approaches do not use the two most difficult landmarks, each located in one earlobe (see 19 landmarks results in Table 4). We have also evaluated 3DDE without the two earlobe landmarks. In this case we get an NME of 2.01, the best reported result.

Finally, we have also evaluated 3DDE with the newly released WFLW data set (Wu et al., 2018). In enables us to evaluate different sources of variability (*i.e.*, expressions, illumination, make-up, occlusions and blur). In Table 5 we provide the results of various competing methods (Wu et al., 2018), normalized by the eye corners distance. 3DDE outperforms its

Method	19 landmarks height	21 landmarks height
	NME	NME
PIFAS (Jourabloo et al., 2017)	-	4.45
CFSS (Zhu et al., 2015)	3.92	-
CCL (Zhu et al., 2016)	2.72	-
DAC-CSR (Feng et al., 2017)	2.27	-
Binary-CNN (Bulat and Tzimiropoulos, 2017)	-	2.85
PCD-CNN (Kumar and Chellappa, 2018)	-	2.40
TSR (Lv et al., 2017)	2.17	-
DCFE (Valle et al., 2018)	2.12	2.17
3DDE	**2.01**	**2.06**

Table 4. Error of face alignment methods on AFLW.

competitors in all the WFLW subsets by a large margin. We hypothesize that the reason for this is that the hybrid approach in 3DDE can be trained with less samples that some of its most prominent competitors and at the same time provide a very accurate face shape (see Fig. 5). Moreover, we achieve the best *AUC* in all subsets, which determines that 3DDE is the best approach under all capture conditions (easy/frontal and difficult/profile) including all subsets that contain several types of difficulties.

4.5. Ablation study

3DDE is based on three key ideas: 3D initialization, a cascaded ERT regressor operating on probabilistic CNN features and a coarse-to-fine scheme. In this section we analyze the contribution of each one to the overall performance of our algorithm.

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}
ESR (Cao et al., 2012)	11.13	27.74	35.24	25.88	1.77	90.18	11.47	19.81	42.04	10.49	29.53	30.80	11.05	24.85	38.84	13.75	19.46	47.28	12.20	22.04	41.40
SDM (Xiong and la Torre, 2013)	10.29	30.02	29.40	24.10	2.26	84.36	11.45	22.93	33.44	9.32	32.37	26.22	9.38	31.25	27.67	13.03	20.60	41.85	11.28	23.98	35.32
CFSS (Zhu et al., 2015)	9.07	36.59	20.56	21.36	6.32	66.26	10.09	31.57	23.25	8.30	38.54	17.34	8.74	36.91	21.84	11.76	26.88	32.88	9.96	30.37	23.67
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
3DDE	**4.68**	**55.44**	**5.04**	**8.62**	**26.40**	**22.39**	**5.21**	**51.75**	**5.41**	**4.65**	**56.02**	**3.86**	**4.60**	**55.36**	**6.79**	**5.77**	**46.92**	**9.37**	**5.41**	**49.57**	**6.72**

Table 5. Error of face alignment methods on WFLW.

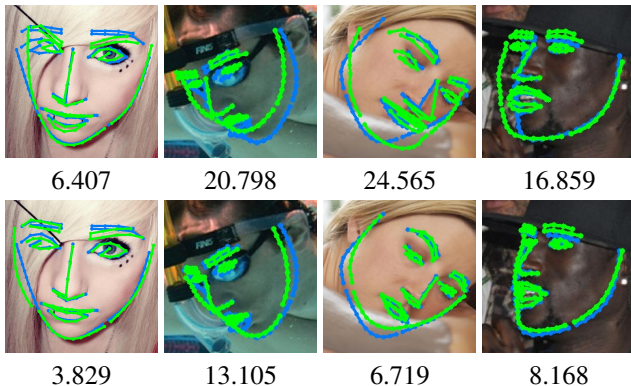| 6.407 | 20.798 | 24.565 | 16.859 |
| 3.829 | 13.105 | 6.719 | 8.168 |

Fig. 5. First row shows LAB (Wu et al., 2018) results, second row 3DDE results. We report the corresponding NME normalized by the eye corners distance. Blue and green colours represent ground truth and predictions respectively.

In Table 6 we show the results obtained by different configurations of our framework when evaluated on WFLW. We have chosen WFLW in our study because it allows the analysis of results stratified by different types of difficulties (*i.e.*, facial expressions, large poses, illumination changes, etc.). In this case, since there are many profile faces, we use the height as normalization for the NME. So, the numerical values are not directly comparable to those in Table 5. MS stands for "mean shape initialization" of the ERT. 3D means to initialize the ERT with the procedure in section 3.1. SE denotes using plain gray level features for the ERT whereas DE denotes using probability maps produced by the CNN to train the ERT. Finally CF stands for using the coarse-to-fine scheme.

When combined with the cascaded ERT, the 3D initialization is key to achieve top overall performance, see CNN+MS+DE vs CNN+3D+DE in the *full* subset. The reason for this is that, in the 3D case, the initialization takes care of the rigid component

of face pose so that the ERT cascade only models non-rigid deformations. Moreover, the projection of the 3D face model is a correct 2D shape, a requirement for the ERT to converge to a valid face shape (Cao et al., 2014). Of course, the 3D initialization is fundamental to achieve good performance in presence of large face rotations. So, it provides the largest improvement in the *pose* subset.

The use of CNN probability maps improves the NME in the *full* data set in about 20% (see CNN+3D+SE vs CNN+3D+DE). The large receptive fields of CNNs are specially helpful in challenging situations, specifically those in the *pose* and *occlusion* subsets.

The coarse-to-fine strategy in our cascaded ERT provides significative local improvements in difficult cases, with rare facial part combinations (see Fig. 6a). For this reason, the largest gain of CNN+3D+DE+CF vs CNN+3D+DE occurs in the *expressions* subset. Although this strategy provides improvements in all the data base subsets, the actual NME differences are washed out when averaged over the number of landmarks in the face and the number of images in the subset. They may be appreciated by looking into specific data subsets or samples (see Fig. 6a), such as the left eyebrow/eye location improvement in Fig. 6b and 6c (best viewed after zoom-in).

Finally, we analyze the NME distribution produced by the rigid initialization and the final 3DDE model (see Fig. 7). Using the model trained for the WFLW experiment, we align the 2500 test samples of WFLW and plot the distribution of NMEs, produced both with the CNN+3D regressor (softPOSIT result) and the full CNN+3D+DE+CF regressor (3DDE result). The val-

Method	Full height			Pose height			Expression height			Illumination height			Make-up height			Occlusion height			Blur height		
	NME	AUC_4	FR_4	NME	AUC_4	FR_4	NME	AUC_4	FR_4	NME	AUC_4	FR_4	NME	AUC_4	FR_4	NME	AUC_4	FR_4	NME	AUC_4	FR_4
CNN+3D+SE	2.52	41.10	11.56	3.53	24.08	28.83	2.90	33.22	15.92	2.53	41.85	10.45	2.59	39.08	15.53	3.06	31.10	22.14	2.91	33.98	15.78
CNN+MS+DE	2.23	49.77	7.04	3.33	35.13	17.79	2.56	45.15	8.91	2.17	49.29	5.87	2.33	46.85	9.70	2.69	40.33	12.90	2.53	42.71	9.57
CNN+3D+DE	2.03	51.14	5.47	2.68	39.55	11.96	2.21	46.66	7.96	2.11	50.09	5.01	2.13	48.57	7.28	2.56	40.83	12.36	2.40	43.84	8.27
CNN+3D+DE+CF	2.01	51.67	5.20	2.63	39.90	10.73	2.15	48.19	5.73	2.06	50.79	4.87	2.12	49.05	7.28	2.54	40.94	12.22	2.39	43.93	8.02

Table 6. Ablation study. MS and 3D represent the 2D mean shape and 2D projections of the 3D mean face respectively. SE and DE represent the type of features used in the cascade being simple grayscale features and deep probability maps features respectively. The CNN+3D+DE+CF row represents the full 3DDE approach results.
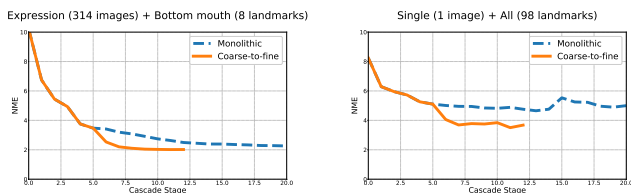
(a) Evolution of the NME through the different stages in the cascade

(b) Monolithic   (c) Coarse-to-fine

Fig. 6. Example of a monolithic ERT regressor vs. our coarse-to-fine approach. (a) NME evolution through the stages in the cascade (left plot, 8 mouth landmarks for all test images in the expression subset; right plot, all 98 landmarks in one image). (b) predicted shape and zoom-in with a monolithic regressor. (c) predicted shape and zoom-in with our coarse-to-fine approach.

Fig. 7. Sample distribution of NMEs produced by the CNN+3D and 3DDE regressors. We use the height as normalization for the NME.

ues of percentiles 10 and 90 of the NME distribution are 3.71 and 6.87 for the `CNN+3D` regressor and 1.03 and 3.32 for the `CNN+3D+DE+CF` one. So, on average, the full regressor reduces in about 60% the NME achieved by the rigid initialization.

4.6. Cross-dataset evaluation

In this section we perform cross-dataset experiments to evaluate the quality of present benchmarks and the generalization of the regressors trained on them. Here we benefit from the fact that 3DDE may be trained in a semi-supervised way, *i.e.*, using data sets with missing or unlabeled landmarks. To this end we select 24 distinct facial landmarks (see Fig. 8a). We consider them distinct because they may be accurately located by
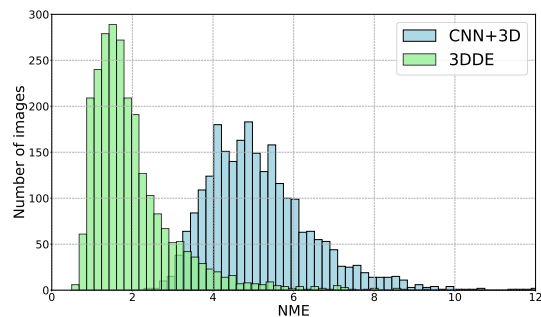
a human annotator. We train and evaluate 3DDE respectively with the training and test sets of each data base. We have also performed one more experiment training 3DDE with the training sets of all data bases and evaluating it successively with the tests sets of each of them, we denote this experiment with label `All`.

In Table 7 we show the results of our evaluation. The smallest data base, COFW, has the worst cross-dataset results. On the other hand, the data set with greatest diversity, WFLW, has the best results. Moreover, the model `All`, trained with the training sets of all data bases, is able to improve, in all cross-dataset experiments, the models trained in a single data set. However, the most prominent outcome of this experiment is that we always achieve the best result when training with the train subset of the same data base. This holds even when compared against the model trained with all data sets, confirming the existence of the so-called "data set bias" in current benchmarks (Torralba and Efros, 2011).

Test \ Train	300W	COFW	AFLW	WFLW	All
300W	**2.00**	3.11	4.90	3.44	4.15
COFW	3.68	**2.09**	4.56	4.03	4.19
AFLW	4.19	2.51	**2.15**	3.29	2.65
WFLW	2.57	2.53	3.28	**1.70**	2.71
All	2.34	2.23	2.41	1.96	**2.26**

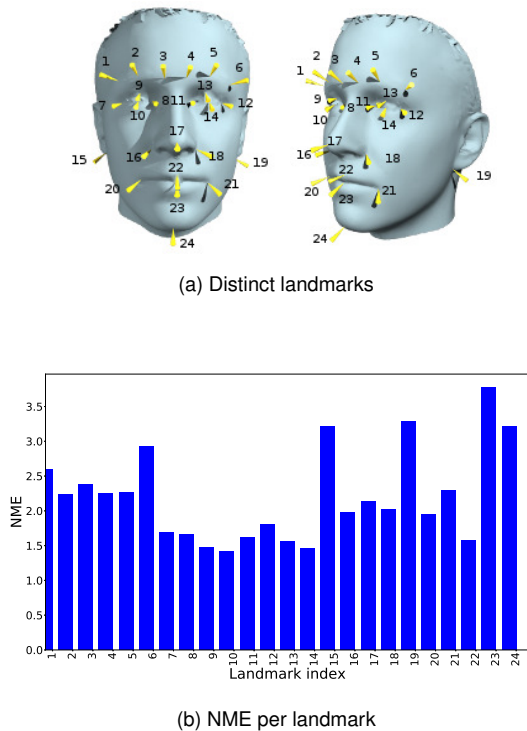**Table 7. Cross-dataset experiment using only distinct landmarks to compute NME normalized by height.**

(a) Distinct landmarks

(b) NME per landmark

Fig. 8. Location of distinct face landmarks and the NME related to each landmark.

In a final experiment we use model `All` to evaluate the NME of each landmark using the test sets of all data sets (see Fig. 8b). The landmarks with highest NME are those related to the ears, the bottom of the mouth and the chin.

5. Conclusions

We have introduced 3DDE, a robust face alignment method that leverages on good properties of CNNs, cascade of ERT and 3D face models. The CNN provides robust landmark estimations with weak face shape enforcement. The ERT is able to enforce the face shape and achieve better accuracy in landmark detection, but it only converges with a good initialization. Finally, 3D models exploit face orientation information to improve self-occlusion estimation.

3DDE is initialized by robustly fitting a 3D face model to the probability maps produced by the CNN. The 3D model enables 3DDE to handle self-occlusions and successfully deal with both frontal and profile faces. Once initialized, the cascade of ERT only models the non-rigid component of face motion. It provides various benefits, namely, it enforces shape consistency, may be trained with unlabeled landmarks, estimate landmark visibility and efficiently parallelize the execution of the regression trees within each stage. We have additionally introduced a coarse-to-fine scheme within the cascade of 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. This is a fundamental limitation of implicit shape models addressed by 3DDE.

In the experiments we have shown that 3DDE improves, as far as we know, the state-of-the-art performance in 300W, COFW, AFLW and WFLW data sets. In our ablation analysis we have shown that all the components of the system critically contribute to the final result.

The availability of large annotated data sets has encouraged research in this area with important performance improvements in recent years. However, as shown in Fig. 9, this problem is still far from being completely solved. A critical question here is whether the models trained with present data sets will generalize to the situations present in real-life operation. The cross-dataset experiments performed reveal the existence of a significant data set bias in present benchmarks that limit the generalization of models trained with them. So, further work in this direction is required to improve the performance of present face alignment algorithms.
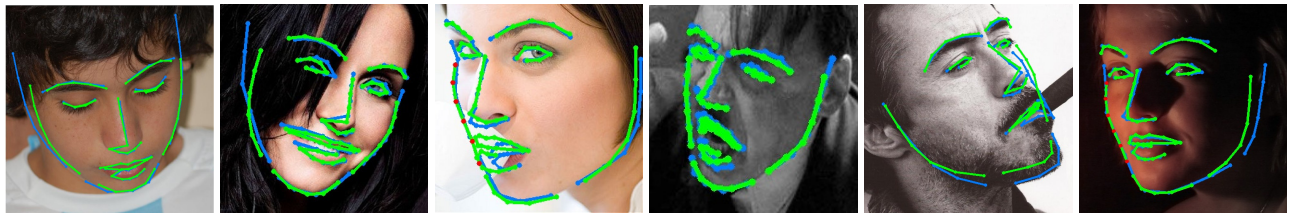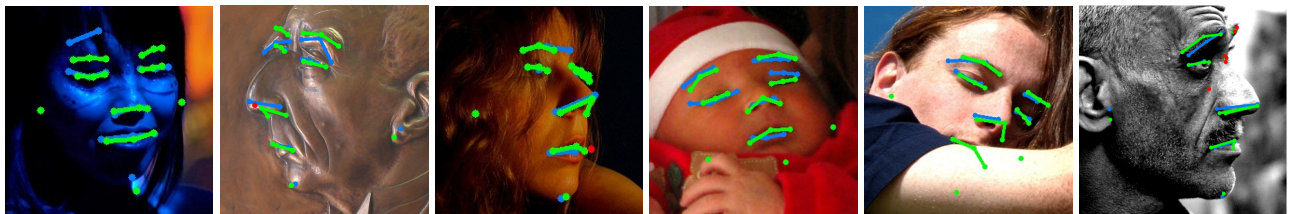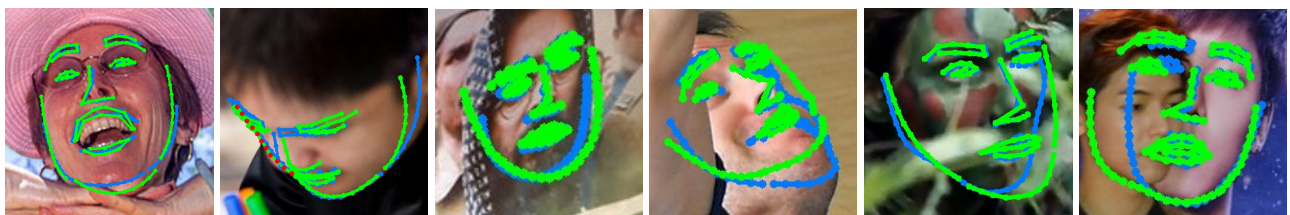
(a) 300W public

(b) 300W private

(c) COFW

(d) AFLW

(e) WFLW

Fig. 9. Representative results considered errors using 3DDE in 300W, COFW, AFLW and WFLW testing subsets. Blue colour represents ground truth, green and red colours point out visible and non-visible shape predictions respectively.

anonymous reviewers for their comments.

References

Alahi, A., Ortiz, R., Vandergheynst, P., 2012. FREAK: fast retina keypoint, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 510–517.

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

Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N., 2011. Localizing parts of faces using a consensus of exemplars, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 545–552.

Blanz, V., Vetter, T., 2003. Face recognition based on fitting a 3D morphable model. IEEE Trans. on Pattern Analysis and Machine Intelligence , 1063–1074.

Bulat, A., Tzimiropoulos, G., 2017. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, in: Proc. International Conference on Computer Vision, pp. 3726–3734.

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.

Cao, X., Wei, Y., Wen, F., Sun, J., 2012. Face alignment by explicit shape regression, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2887–2894.

Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. International Journal of Computer Vision 107, 177–190.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 1998. Active appearance models, in: Proc. European Conference on Computer Vision, pp. 484–498.

Dantone, M., Gall, J., Fanelli, G., Gool, L.V., 2012. Real-time facial feature detection using conditional regression forests, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2578–2585.

David, P., DeMenthon, D., Duraiswami, R., Samet, H., 2004. Softposit: Simultaneous pose and correspondence determination. International Journal of Computer Vision 59, 259–284.

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

Feng, Z., Kittler, J., Christmas, W.J., Huber, P., Wu, X., 2017. 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, pp. 3681–3690.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning. Springer.

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.

Jin, X., Tan, X., 2017. Face alignment in-the-wild: A survey. Computer Vision and Image Understanding 162, 1–22.

Jourabloo, A., Ye, M., Liu, X., Ren, L., 2017. Pose-invariant face alignment with a single CNN, in: Proc. International Conference on Computer Vision, pp. 3219–3228.

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874.

Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H., 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in: IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, pp. 2144–2151.

Kowalski, M., Naruniec, J., 2016. Face alignment using k-cluster regression forests with weighted splitting. IEEE Signal Processing Letters 23, 1567–1571.

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.

Kumar, A., Chellappa, R., 2018. Disentangling 3d pose in a dendritic CNN for unconstrained 2d face alignment, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 430–439.

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.

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.

Martinez, A.M., Du, S., 2012. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. Journal of Machine Learning Research 13, 1589–1608.

Ren, S., Cao, X., Wei, Y., Sun, J., 2016. Face alignment via regressing local binary features. IEEE Trans. on Image Processing 25, 1233–1245.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention - MICCAI, pp. 234–241.

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.

Shen, W., Deng, K., Bai, X., Leyvand, T., Guo, B., Tu, Z., 2014. Exemplar-based human action pose correction. IEEE Trans. Cybernetics 44, 1053–1066.

Soltanpour, S., Boufama, B., Wu, Q.M.J., 2017. A survey of local feature methods for 3D face recognition. Pattern Recognition 72, 391–406.

Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483.

Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1521–1528.

Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S., 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4177–4187.

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., 2015. Robust facial landmark detection under significant head poses and occlusion, in: Proc. International Conference on Computer Vision, pp. 234–241.

Xiao, S., Feng, J., Liu, L., Nie, X., Wang, W., Yan, S., Kassim, A.A., 2017. Recurrent 3D-2D dual learning for large-pose facial landmark detection, in: Proc. International Conference on Computer Vision, pp. 1642–1651.

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.

Xiong, X., la Torre, F.D., 2013. Supervised descent method and its applications to face alignment, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539.

Xiong, X., la Torre, F.D., 2015. Global supervised descent method, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2664–2673.

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.

Yu, X., Zhou, F., Chandraker, M., 2016. Deep deformation network for object landmark localization, in: Proc. European Conference on Computer Vision, pp. 52–70.

Zhang, J., Hu, H., 2018. Exemplar-based cascaded stacked auto-encoder networks for robust face alignment. Computer Vision and Image Understanding 171, 95–103.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2014. Facial landmark detection by deep multi-task learning, in: Proc. European Conference on Computer Vision, pp. 94–108.

Zhao, R., Wang, Y., Benitez-Quiroz, C.F., Liu, Y., Martínez, A.M., 2016. Fast and precise face alignment and 3D shape reconstruction from a single 2D image, in: Proc. European Conference on Computer Vision Workshops, pp. 590–603.

Zhu, S., Li, C., Change, C., Tang, X., 2016. Unconstrained face alignment via cascaded compositional learning, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3409–3417.

Zhu, S., Li, C., Loy, C.C., Tang, X., 2015. Face alignment by coarse-to-fine shape searching, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4998–5006.

Zhu, X., Liu, X., Lei, Z., Li, S.Z., 2017. Face alignment in full pose range: A 3D total solution. IEEE Trans. on Pattern Analysis and Machine Intelligence 41, 78–92.