# $L_{2,1}$-based regression and prediction accumulation across views for robust facial landmark detection

Brais Martinez[1]

*School of Computer Science, University of Nottingham, U.K.*

Michel F. Valstar

*School of Computer Science, University of Nottingham, U.K.*

**Abstract**

We propose a new methodology for facial landmark detection. Similar to other state-of-the-art methods, we rely on the use of cascaded regression to perform inference, and we use a feature representation that results from concatenating 66 HOG descriptors, one per landmark. However, we propose a novel regression method that substitutes the commonly used Least Squares regressor. This new method makes use of the $L_{2,1}$ norm, and it is designed to increase the robustness of the regressor to poor initialisations (e.g., due to large out of plane head poses) or partial occlusions. Furthermore, we propose to use multiple initialisations, consisting of both spatial translation and 4 head poses corresponding to different pan rotations. These estimates are aggregated into a single prediction in a robust manner. Both strategies are designed to improve the convergence behaviour of the algorithm, so that it can cope with the challenges of in-the-wild data. We further detail some important experimental details, and show extensive performance comparisons highlighting the performance improvement attained by the method proposed here.

*Keywords:* Facial landmark detection; regression; 300W challenge

*Corresponding author. Tel. +44 (0) 115 951 4251; Fax: +44 (0) 115 951 4254.
E-mail address: brais.martinez@nottingham.ac.uk (B. Martinez), michel.valstar@nottingham.ac.uk (M.F. Valstar)

## 1. Introduction

Existing works on facial landmark detection are often divided into holistic models (e.g. AAM [1, 2, 3]), and part-based models. Traditionally, part-based models iteratively alternate between two steps: the construction of landmark-specific response maps, and the shape fitting step. The response map construction relies on the use of landmark-specific classifiers trained to fire when evaluated at the correct landmark location. A response map for a landmark is constructed by scanning a classifier with a probabilistic output over a region of interest in a sliding window manner [4]. The subsequent shape fitting step consists of finding the landmark locations maximising individual responses, but constrained to having a valid shape according to the shape model (most typically a Point Distribution Model [5]).

The two most challenging aspects of part-based classifier models are (1) training classifiers that are sensitive enough to perform fine grained detection, and (2) most importantly, the extreme challenge of the shape fitting stage. The latter process is plagued with local minima and often results in a costly maximisation procedure. The most notable efforts within this group are those of Belhumeur et al. [6], and the DRMF [7]. The former used a RANSAC-type shape fitting, while the latter used a discriminative regression-based model predicting shape increments. However, obtaining reliable performance using these approaches implies a strong implementation effort and significant know-how and, even then, their performance now trails behind that of other state-of-the-art methods.

An important exception both in terms of the theoretical framework and the practical performance is that of Zhu and Ramanan [29]. In this work, the authors used a discriminative classifier and part-based model consisting on an adaptation of the successful Deformable Parts Model [23] for facial landmarking. The main difference arises from the use of a tree-based graphical model to capture the face shape. Exact inference becomes possible, but multiple pose-wise experts can be used to capture different head poses, including profile faces.

While their ability to perform exact inference is remarkable and very useful in practise, their precision is lower than for other methods (provided they converge), and detection can be slow despite the strong speed-up provided by a complex yet efficient implementation.

Alternatively, Valstar et al. [8] proposed to drive the local search by regressors performing direct displacement prediction instead of by classifiers measuring landmark fitness. Landmark-specific regression models were trained to this end, with each regressor being tasked with predicting the displacements in the $x$ and $y$ direction from the test location directly to the true target location. While this resulted in promising performance, this approach still has several shortcomings, such as its lack of robustness to erroneous regressor predictions, or the effective inclusion of shape constraints, in particular for non-frontal head poses. Further improvements on regression-based landmarking was attained by combining multiple regression predictions into the equivalent of response maps [9], [10]. Thus, while the response maps obtained were typically more precise than those obtained with classification approaches, the shape alignment step was still hindering practical performance.

A new breakthrough was proposed by Cao et al.[11]. Firstly, they adopted the cascaded regression framework of Dollár et al. [12], which powered regression-based predictions to allow for inference being simultaneously robust and precise. Secondly, they proposed to directly estimate the shape increments as a whole. That is to say, instead of having a per-landmark model, they used a combined model, taking the whole face appearance as input, and predicting increments for the whole shape. This allowed bypassing the cumbersome shape fitting step, and shape consistency was enforced through the joint prediction. It is interesting to note that face shapes are assumed to lie in a linear subspace (once rigid parameters are eliminated).

However, this approach really became the state-of-the-art due to the work of Xiong & De la Torre [13]. While the authors followed a similar approach to that of Cao et al. [11], they managed to greatly simplify the methodology by adopting HOG features and only relying on least squares for inference. The

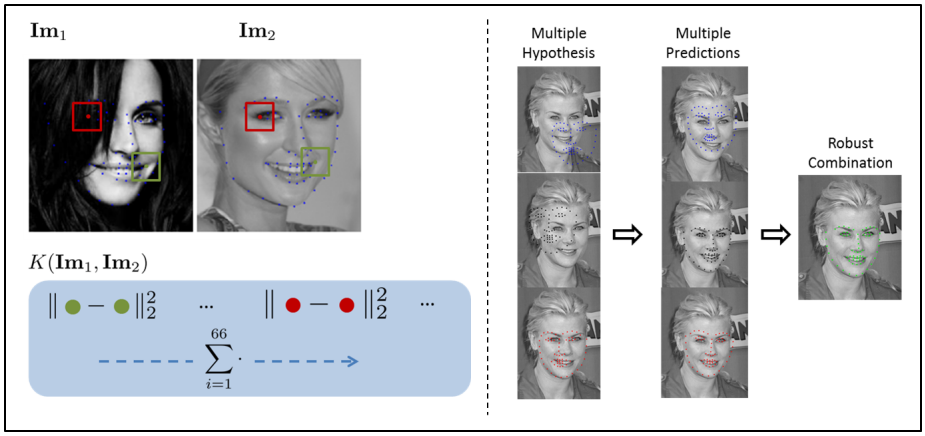
resulting algorithm attained state-of-the-art performance using only 4 matrix multiplications and ran in real time on a standard PC with minimal implementation efforts. The authors also provided an implementation of the method, including extremely well optimised pre-trained models. Furthermore, due to its simplicity, the method can be re-implemented from scratch very easily. Despite its huge advantages, the method of Xiong & De la Torre still presents some drawbacks. Firstly, there is no confidence score for each prediction step, so that there is no knowledge of whether the inference actually improved the solution. Thus, it is not possible to use multiple initialisations or mixture models, which is particularly important for largely non-frontal head poses. Secondly, the use of least squares is not robust and it is thus not ideal in the presence of partial occlusions or when a subset of the landmarks are far off from their ground truth location.

Many works have since then built upon the works of Cao et al. [11] and Xiong & De la Torre [13] in different ways. For example, Ren et al. [14] and Kazemi & Sullivan[15] presented extremely fast face alignment algorithms using variants of these ideas. Several works have proposed methods for improving the robustness to partial occlusions. Specifically, Burgos-Artizzu et al. [16] proposed to train a model tasked with detection occlusions explicitly in a discriminative manner. An alternative approach was proposed in Xing et al. [17], where a sparse dictionary learning approach was followed as an alternative to the least squares regression of [13]. This thus constitutes a generative approach rather than discriminative. A specific mechanism within the construction of the dictionary was also included to tackle fitting under partial occlusions. An alternative generative variant of [13] was proposed by Tzimiropoulos [18]. It maintained the PCA-based model traditional for generative models (see e.g. [1, 2]), but as novel elements it used a cascade regression approach and a novel mechanism for removing appearance variation in successive levels of the cascade. The work by Sun et al. [19] proposed instead to use a Convolutional Neural Network approach to model the inference problem at each of the cascade levels. Finally, Yan et al. [20] proposed to use a discriminative ranking model capable

of selecting and combining multiple predictions, each one obtained using the SDM method and using a different initial shape hypothesis. In fact, this last work won the first 300W facial landmark challenge [21].

In this work we build on the previous efforts mentioned above, aiming to tackle the problems of a lack of confidence measures of the predictions and the problem of least squares fragility inherent to [13]. The main methodological contributions of this paper are as follows: Firstly, we propose a new robust regression methodology based on the use the $L_{2,1}$ norm [22]. This norm allows us to compare two shapes in a robust manner, so that sparse error patterns are primed. The details of this approach will be described in 2. Since the resulting distance is not linear, we resort to its kernelisation, and then employ a standard Support Vector Regression technique for inference. Secondly, we resort to multiple initialisations, and employ an estimate aggregation technique in order to combine the resulting estimates in a robust manner [9]. The aim of this process is to increase the robustness to large out-of-plane rotations. In particular, we use four shapes covering a range of pan head rotations, and for each head pose we create a number of initialisations by simply displacing the viewpoint-specific mean shape in a grid manner on the $x$ and $y$ axis. This process is explained in detail in section 3. A depiction of the detection process is summarised in Fig. 1.

While these are the two major methodological components of our method, we have performed other optimisations worth mentioning. Firstly, we use a face detector trained using the Deformable Parts Model [23]. This greatly improves both the precision and the robustness of the initial estimate respect to that of a Viola and Jones face detector. Secondly, the features we use to represent local patches result from first computing a HOG descriptor, and then computing PCA over them [23]. This serves a twofold purpose: it improves the speed of the inference evaluations and increases the precision of the predictions. These and other minor details and aspects of the algorithm will be detailed in Sec. 4.




(a) Comparing two examples using the proposed $L_{2,1}$-based norm.

(b) Multiple predictions from different hypothesis are combined at test time.

Figure 1: Overview of the method. The left image depicts the way the $L_{2,1}$ norm is used to define a kernelised distance between examples, which is integrated into the inference model. At test time, multiple initial hypothesis are considered and the resulting predictions are combined. This is depicted in the right image.

## 2. $L_{2,1}$ norm cascaded regression

One of the remarkable aspects of the work presented by Xiong & De la Torre [13] is the excellent performance attained even when using Least Squares regression, a very simple machine learning method. Much of the excellent performance is due to the use of cascaded regression. We review the principle of cascaded regression in Sec. 2.1, both for completeness, and to define notation. Our first contribution is to change the inference algorithm used in the regression cascade of [13], substituting the Least Squares regression for a novel $L_{2,1}$ norm-based approach. This change is motivated and detailed in Sec. 2.2.

### *2.1. Cascade of linear regressors*

*Inference.* A shape contains $N_{pts}$ landmarks (66 in our case), and it is represented as a $2N_{pts}$-dimensional vector. Inference starts with an initial shape estimate, say $\mathbf{s}^0$, typically given by the face detector[1]. The appearance corresponding to a shape $\mathbf{s}$ is constructed by computing a descriptor (HOG in this case) on a small patch centred at each of the $N_{pts}$ landmarks defined by shape $\mathbf{s}$. The resulting descriptors are then concatenated into a single vector. We use the notation $f(\mathbf{s}, \mathbf{I})$ to indicate that the appearance descriptor is computed for shape $\mathbf{s}$ on image $\mathbf{I}$.

Inference is attained by sequentially applying a set of linear regressors, so that the output of the previous regressor is the input to the next regressor. Specifically, each such linear regressor is defined in [13] as $\{\mathbf{W}^k, \mathbf{b}^k\}_{k=1:N_{it}}$, where $\mathbf{W}^k$ is a matrix containing the regression coefficients, $\mathbf{b}^k$ is the bias term[2] and $N_{it}$ is the total number of iterations in the cascade. $N_{it}$ is fixed, and there is no convergence criterion, so that the chain of regressors is applied in full every time. Specifically, an iteration of the algorithm proceeds as follows:

[1] Bold lower-case letters indicate vectors. All vectors are column vectors unless indicated otherwise. Matrices are typeset as upper-case bold letters. All other letters are scalars.

[2] It is common to simplify the notation by including the bias term within the matrix $\mathbf{W}^k$ and appending a 1 at the end of the input feature vector.

$$\mathbf{x}^k = f(\mathbf{s}^{k-1}, \mathbf{I}) \tag{1}$$

$$\mathbf{y}^k = (\mathbf{W}^k)^T \mathbf{x}^k + \mathbf{b}^k \tag{2}$$

$$\mathbf{s}^k = \mathbf{s}^{k-1} + \mathbf{y}^k \tag{3}$$

where, $\mathbf{I}$ is the test image, and $\mathbf{s}^{N_{it}}$ is the final shape estimate.

*Learning.* We note the images within the training set as $\{\mathbf{I}_j\}_{j=1:N_{im}}$. For each of these images, a set of initial shapes are used $\{\mathbf{s}^0_{i,j}\}_{i=1:N_{init}}$. These multiple initialisations can be obtained by, for example, first registering the mean shape to the ground truth using scaling and translation only, and then perturbing the resulting shape. However, other strategies to generate the initial shapes exist [12], [13].

The first training set is defined as:

$$\begin{aligned} \{(\mathbf{x}^1_{i,j}, \mathbf{y}^1_{i,j})\}_{i=1:N_{init}, j=1:N_{im}} \\ \mathbf{x}^1_{i,j} = f\left(\mathbf{s}^0_{i,j}, \mathbf{I}_j\right) \\ \mathbf{y}^1_{i,j} = \mathbf{s}^*_j - \mathbf{s}^0_{i,j} \end{aligned} \tag{4}$$

where $\mathbf{s}^*_j$ is the ground truth shape for image $j$.

Then, the first regressor can be learnt as:

$$\arg\min_{\mathbf{W}^1, \mathbf{b}^1} \sum_{i=1}^{N_{init}} \sum_{j=1}^{N_{im}} \|\mathbf{s}^*_j - \mathbf{s}^0_{i,j} - {\mathbf{W}^1}^T \mathbf{x}^1_{i,j} - \mathbf{b}^1\| \tag{5}$$

In the general case,

$$\mathbf{s}^k_{i,j} = \mathbf{s}^{k-1}_{i,j} + {\mathbf{W}^k}^T \mathbf{x}^k_{i,j} + \mathbf{b}^k \tag{6}$$

and $\mathbf{W}^k, \mathbf{b}^k$ are obtained using $\mathbf{s}^k_{i,j}$ in a similar manner as in equation 4 and 5.

### 2.2. $L_{2,1}$ norm regression

Our approach follows the same cascaded regression scheme, but we modify the regressor of choice. That is, instead of computing a linear regressor $\left(\mathbf{W}^k, \mathbf{b}^k\right)$ at every step, we compute a non-linear regressor $\mathcal{G}(-;\theta_k)$. The proposed regressor is based on the use of the $L_{2,1}$ norm [22]. Specifically, we want to find a way to compare two feature vectors, say $\mathbf{x}_1$ and $\mathbf{x}_2$, in a robust manner.

Remember that each feature vector $\mathbf{x}$ was generated by computing landmark-specific feature vectors and then concatenating them together into a single vector. We re-define the appearance feature vector, now denoted as $\mathbf{X}$, as the $n \times N_{pts}$ matrix that results from re-ordering the $n$-dimensional per-landmark appearance feature vectors corresponding to the $N_{pts}$ landmarks. That is, instead of concatenating the per-landmark appearance feature descriptors vertically, we concatenate them horizontally, resulting in a matrix rather than a vector. Then, we define the distance between two appearance feature vectors as:

$$d\left(\mathbf{X}_1, \mathbf{X}_2\right) = \|\mathbf{X}_1 - \mathbf{X}_2\|_{2,1} \tag{7}$$

where

$$\|\mathbf{X}\|_{2,1} = \sum_{j=1:N_{pts}} \|\mathbf{X}_{:,j}\|_2 \tag{8}$$

where $\mathbf{X}_{:,j}$ indicates the column $j$ of matrix $\mathbf{X}$.

In doing so, the comparison between two shapes is obtained by first computing the $L_2$ distance between per-landmark representations, obtaining a 66-dimensional vector, and then computing the $L_1$ norm over the resulting vector. It is interesting to note that the (squared) Euclidean distance used in Least Squares regression would result from simply computing the $L_2$ norm again on the 66-dimensional per-landmark $L_2$ distance. However, by substituting the computation of the $L_1$ norm for the $L_2$ norm in the second step, we enforce sparse landmark-to-landmark error patterns. These error patterns are typical in the presence of partial occlusions, so that the occluded landmarks will yield

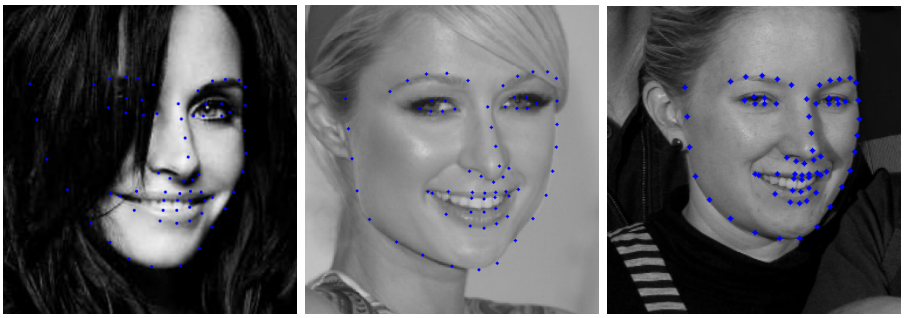


Figure 2: The error measure used should consider the appearance corresponding to the left-hand side shape to be similar to those of the centre and right-hand side images. To this end, it is necessary to deal with sparse landmark-wise error patterns

high $L_2$ errors while the rest of the landmarks will result in low ones. A similar effect happens when there is a large head pose variation between two examples (e.g. a frontal shape is used to initialise the search for a non-frontal head pose), or when contour landmarks are poorly aligned so that the corresponding appearance patterns can be extracted from the background. This is illustrated in Fig. 2. This figure shows the test image (left-hand side) with its ground truth, and two training images with their ground truth shapes. Since the shapes are very similar, we would like to use a distance that considers the associated appearance patterns to be similar. However, the partial occlusion on the test image requires a robust comparison.

The regression function is now non-linear. Thus, we resort to the use of Support Vector Regression (SVR) and use a kernelised version of this norm. Specifically, we compute:

$$\mathcal{K}(\mathbf{X}_1, \mathbf{X}_2) = e^{-\gamma \|\mathbf{X}_1 - \mathbf{X}_2\|_{2,1}} \tag{9}$$

We use an off-the-shelf solver for this problem [24].

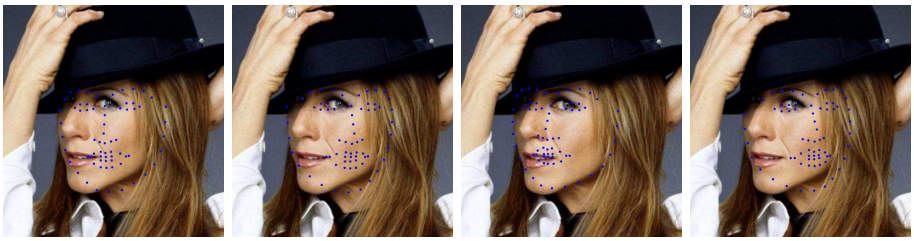


Figure 3: The four initial shapes on an example image (taken form the iBug dataset). Left-hand side examples are of near-frontal head poses, while right-hand side ones correspond to larger head pose rotations.

## 3. Estimate Aggregation

While the use of a robust regressor improves the algorithm performance in images with non-frontal head poses, we further combine this strategy with a multiple initialisation and aggregation strategy [9]. Specifically, for each image, we consider a set of four initial shapes corresponding to distinct head poses, noted $\{\mathbf{s}_i^0\}_{k=1:4}$ (see Fig. Fig. 3). The specific four initial shapes used in here result from the face detection algorithm used ([25], see Sec. 4 for details on the face detector used). In particular, face detection results from applying 4 pose-wise experts, and the pose-wise expert yielding the best score is responsible for the face detector. A per-expert mean shape is constructed by using the subset of all the training faces for which the specific expert provided the detection. This is however a heuristic rule, although in our case this yields better overall performance compared to manually defining the initial shapes to be equally spaced in terms of their rotation angles.

These shapes are fitted to the test image using the bounding box resulting from the face detection process. Then, each shape is perturbed at regular intervals along the $x$ and $y$ axis in a grid-like manner. This can be done for example defining a vector of displacements $\mathbf{v} = (-R, \ldots, -r, 0, r, \ldots R)$, where $r$ is the stride or step size, and $R$ is the maximum displacement. Let us define $\Delta\mathbf{x}_i$ as a $2N_{pts}$-dimensional vector with $\mathbf{v}(i)$ in its first $N_{pts}$ dimensions and 0 on the other dimensions, while $\Delta\mathbf{y}_i$ is defined equivalently but with 0 on the first $N_{pts}$

dimensions instead. We can then define set of initial shapes as:

$$\mathbf{s}^0_{k,i,j} + \Delta\mathbf{x}_i + \Delta\mathbf{y}_j \qquad k = 1:4; i,j = 1:|\mathbf{v}| \tag{10}$$

accounting for a total of $4 \times |\mathbf{v}| \times |\mathbf{v}|$ initial shapes.

The first step of the regression cascade is then computed, in our case using the methodology explained in Section 2.2. This yields a set of predictions $\mathbf{s}^1_{k,i,j}$. Then we aim to combine these estimates into a single prediction. This is done by using a prediction aggregation strategy, in a similar manner to Local Evidence Aggregation [9]. Specifically, we consider a 1-dimensional Gaussian distribution with fixed covariance $\sigma_0$. Then, we define a response map for each landmark $l$, noted $\mathbf{R}^l$, as follows:

$$\mathbf{R}^l(x,y) = \sum_{i,j,k} \mathcal{N}\left(x; \mathbf{s}^1_{k,i,j}(l), \sigma_0\right) \mathcal{N}\left(y; \mathbf{s}^1_{k,i,j}(l+N_{pts}), \sigma_0\right) \tag{11}$$

This process actually performs a Kernel Density Estimation using a Gaussian isotropic kernel over the regressor predictions. Each of the response maps encodes the belief of a certain image location being the true landmark location when considering all the estimates simultaneously. However, when this belief is only considered in a local manner, i.e., if we were to pick the maximum of each response map as the prediction, the resulting shape would not be anthropomorphically consistent. Thus, the aim is now to find the consistent shape that maximises the individual responses:

$$\hat{\mathbf{s}}^1 = \arg\max_{\mathbf{s}} \sum_{l=1}^{N_{pts}} \mathbf{R}^l\left(\mathbf{s}(l), \mathbf{s}(l+N_{pts})\right) \qquad \text{s.t. } \mathbf{s} \text{ is valid} \tag{12}$$

However, this is a very challenging optimisation (in fact, it has been one of the most pressing optimisation problems for facial landmark detection over the last decade). In order to avoid complex procedures at this stage, which is not the main focus of this work, we resort to the simple strategy of restricting the

search space to the estimates $\mathbf{s}^1_{k,i,j}$. That is to say, we define:

$$\hat{\mathbf{s}}^1 = \arg\max_{\mathbf{s}_{i,j,k}} \sum_{l=1}^{N_{pts}} \mathbf{R}^l \left( \mathbf{s}^1_{i,j,k}(l), \mathbf{s}^1_{i,j,k}(l + N_{pts}) \right) \tag{13}$$

This serves the purpose of improving performance in the presence of non-frontal head poses and of less precise face detections (arguably, the precision of the face detection is lower for non-frontal head poses, thus both cases often co-occur). While classification-based approaches can rely on the score of the classifier (e.g. using logistic regression [4]), regression-based approaches do not have an equivalent. Thus, we use the Local Evidence Aggregation property highlighted by Martinez et al. [9], for which the accumulation of regression predictions result from meaningful input patterns. Instead, patterns unseen during training, such as those too far from the ground truth either in terms of the head pose or of the displacement, result in random predictions which do not accumulate.

We repeat this process for the second iteration of the algorithm. However, in this case we do not use 4 pose-wise shapes. Instead, we only consider $\hat{\mathbf{s}}^1$, and perturb it by translating it by a smaller amount than used in the first iteration. The remaining iterations do not include this procedure as it was shown ineffective in these cases. This is not surprising, as the algorithm converges very quickly and the last iterations only fine-tune the prediction.

## 4. Implementation Details

*Features.* HOG features [26] have become one of the standard appearance descriptors for facial landmarking, as they are very suited to in-the-wild landmarking. They are robust to variations in illumination, as they rely on gradients and the histogram representation is normalised to one. In addition, the effect of non-frontal head pose rotations can be locally approximated by an affine transformation, to which HOG features are robust. We follow the same procedure as Felzenszwalb et al. [23] and compute a HOG-PCA descriptor. HOG-PCA computes PCA after computing the HOG descriptor for each landmark across

the entire training set. We optimised the number of components to be retained, and found that optimal performance was attained with as few as 10 dimensions for the first two iterations, and 30 for the next two. It is important to note that the PCA is done per landmark, and thus we use a total of 660 and 1980 features, respectively. The benefits of using HOG-PCA is two-fold. Firstly, we are now using a non-linear SVR and thus the model includes all the support vectors. If the feature dimensionality is very large, then the run-time memory requirements can quickly become prohibitive. Secondly, we experimentally found that the performance improved significantly compared to using the full HOG descriptor. It is also interesting to note that using a single global PCA on the concatenated representation is incompatible with the use of the $L_{2,1}$ norm.

*Face detector.* Due to the frequent presence of non-frontal head poses, partial occlusions and the use of in-the-wild imagery in general, we have opted for using a face detector obtained by training the successful Deformable Parts Model [23] for this specific task [25]. The resulting detection is not only robust to the aforementioned situations, but also offers higher precision in terms of the initial shape estimate. While the face detector is trained with a mixture of four different pose-wise components, we avoid using the head pose corresponding to the component that yielded the detection. We have experimentally found that the component resulting in the detection is not always correct, in particular for non-frontal head poses. Thus, the initial shape would in these cases deviate too much from the true landmark locations to obtain a correct detection. Instead, we only rely on the strategy described in Sec. 3 to overcome the problem of how to initialise non-frontal head poses.

*Internal parameters.* The described algorithm depends on some parameters which need to be optimised. They include the kernel parameter $\gamma$ (see Sec. 2.2), the stride of the perturbation grid $r$, the maximum perturbation $R$, the variance $\sigma_0$ (see Sec. 3), and the number of PCA dimensions of the feature representation.

We defined the parameter $\gamma$ heuristically as follows:

$$\gamma = \frac{1}{\underset{\mathbf{X_i},\mathbf{X_j}}{\text{median}}\{d(\mathbf{X_i},\mathbf{X_j})\} - \underset{\mathbf{X_i},\mathbf{X_j}}{\min}\{d(\mathbf{X_i},\mathbf{X_j})\}} \tag{14}$$

However, there might be space for further performance improvement by fine-tuning this parameter on a validation set.

The remaindert of the parameters were optimised by using LFPW [27] and Helen [28] datasets for training, and the AFW dataset [29] as a validation set. We used several performance measures to decide the best parameters, including the mean inter-ocular distance (iod) normalised error, the median iod-normalised error, the cumulative error curves, and the percentage of images in which the error was reduced respect to the previous estimate. We put particular emphasis in reducing the amount of gross errors on the first iterations, priming robustness over precision (hence the complementary error measures considered and why we decided on the parameters by visual inspection of these values). The resulting parameters were $r = 5$ and $R = 20$ for the first iteration, and $r = 5$ and $R = 15$ for the second one. This yields a total of $4 \times 9 \times 9 = 324$ and $7 \times 7 = 49$ test shapes on each of the two initial iterations respectively. We found the performance to be robust with respect to the value of $\sigma_0$, and we defined it as 0.03 times the length of the (square) face bounding box side. Retaining the first 10 PCA dimensions for the two first iterations of the cascade, and 30 dimensions for the remainder were found to work optimally. We perform 4 iterations of the regression cascade as performance improve marginally to none for the fifth iteration.

*Prediction target.* Similar to [30], we aim to predict the parameters of a shape model rather than the landmark locations. To this end, we employ the 3D Point Distribution Model provided by [4]. However, this shape model contains 66 landmarks. The predictions for the last 2 landmarks are added after the detection of the other 66 is finished. The main reason behind this is to reduce the computational cost and, most importantly, the memory storage requirements. While each output dimension requires its own regressor, we reduce the number

of output dimensions from the original 136 dimensions to 30 (6 dimensions for rigid parameters and 24 for flexible parameters).

*Face registration.* At the beginning of every iteration, we register the current shape estimate to the mean shape using a Procrustes transformation. Then, the same transformation is applied to the image to normalise the face image with respect to head rotation and scaling prior to the feature extraction step. In the first step, we normalise to a mean shape corresponding to a face bounding box size of 100 pixels. For subsequent steps, we use a mean shape corresponding to a face bounding box size of $200 \times 200$ pixels. Registering to a larger size can affect the robustness of the prediction, as the relative distances to the ground truth are increased. Later steps of the cascade relate however to the refinement of the prediction, and it is then useful to be able to use more detailed images.

## 5. Experimental Results

*The data used.* We have trained our model using the training partition of the LFPW [27] dataset, the training partition of the Helen dataset [28], and the AFW dataset [29]. Testing datasets include the testing partition of both LFPW and Helen, the IBUG dataset [31], and the hidden dataset used by the challenge organisers. While all of these datasets contain in-the-wild images, they are of varying difficulty. The LFPW and Helen datasets contain mostly well-illuminated frontal head poses with limited partial occlusions. Thus, they are the easiest of the datasets considered. However, the Helen dataset contains more expressive faces, although the image resolution is also larger in general. The AFW dataset contains in comparison more non-frontal head poses than LFPW and Helen, and has an intermediate difficulty. Finally, the IBUG dataset consists only of 135 images, but it is the most challenging dataset of all. It contains all kind of frequent self-occlusions, largely non-frontal head poses and a large variety of illumination conditions.

*Error measure.* The graphs shown in this article were constructed using the function to compute the error provided by the challenge organisers. Firstly, the

error per image is computed as the Inter-ocular distance (iod) normalised error. To this end, the average per-landmark Euclidean distance between the detected location and the true target location is computed. Then, the resulting value is divided by the Euclidean distance between the two landmarks corresponding to the outer corners of the eye, computed using the ground truth. It is interesting to note that the iod-normalised distance is sometimes defined as the distance between the centre of the eyes, resulting in larger values to the iod-normalised error. We also report both the error for inner-facial landmarks (excluding landmarks lying on the contour of the face), and for all landmarks (including the contour ones). Please note that, since our shape model contains 66 landmarks, the errors reported here are computed over 66 landmarks. The only exception is the challenge results.

*Reproducing the results.* Upon acceptance of this paper, we will provide a publicly available implementation of our method on the authors' websites. The code is exactly the same as that submitted to the 300W challenge, including all the internal parameters, except for the correction of a bug regarding the face detection. Thus, the performance on the 300W challenge data is actually higher than reported in this paper.

While the challenge data was restricted to contain only one face per image, some images on other datasets contain several faces. In these cases, we have manually selected the automatically-detected face bounding box corresponding to the right face (please note that the face detection is still automatic!). We have also corrected some other cases on the IBUG dataset where, while only one face is present in the image, the highest-scored face detection is wrong. This accounted for 10 images out of the 135 contained in the dataset. In these cases, we selected the automatically-detected bounding box better fitting the face. In order to allow the reproduction of the results presented here, we provide the bounding boxes used to generate the graphs. The code then takes the bounding box as an optional input while, if the bounding box is not specified, the face detection routine is then executed to obtain one. In this case, only the highest-

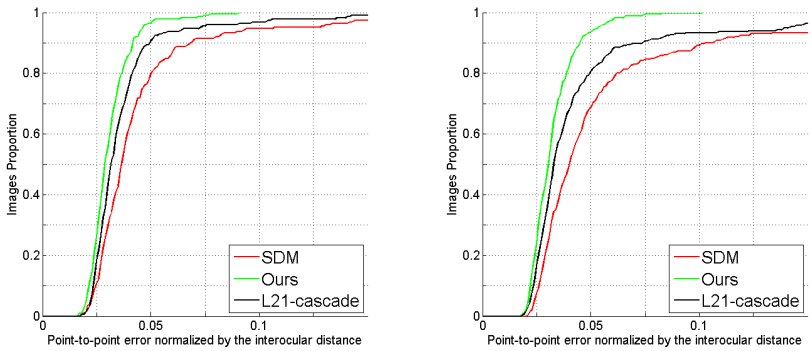
scoring bounding box is considered, which might result in the facial landmarking of the wrong face.

*Relative merit of the algorithm components.* The first set of experiments shown in here is designed to clarify the specific merits of each of the proposed methodological improvements (i.e., the use of the $L_{2,1}$ norm-based regression, and the regression aggregation procedure). To this end, we show the performance of our implementation of the Supervised Descent Method (SDM)[3] [13], a version of the cascaded regression using the $L_{2,1}$ norm-based regression for inference, and finally the performance of the full algorithm. It is important to note that both the proposed regression model and the aggregation strategy are designed to improve the robustness of the method, but whenever there is convergence, they might not result in a more precise fitting than the SDM. Robustness is particularly important on difficult datasets such as the AFW or the IBUG datasets. However, of the two, we only report performance on the IBUG dataset. This is due to the inclusion of examples from the AFW dataset on our training set.

The cumulative error functions for the 49 inner landmarks can be seen on Fig. 4. It is possible to see how the proposed method is characterised by increased robustness, as the largest gains correspond to the larger error values. That is to say, when there is convergence, it might not be more accurate. However, the proportion of images converging to the right solution is boosted. The performance difference becomes abysmal for the IBUG dataset, where most images are very challenging. However, it is also remarkable that for the LFPW and Helen datasets practically all images have an error under 0.1.

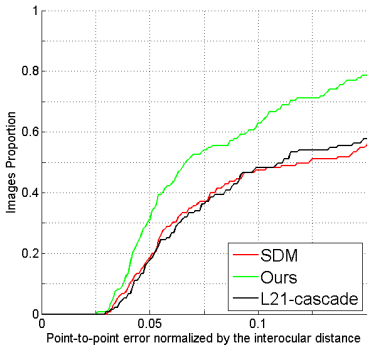
We further show in Table 1 the performance improvement obtained, in terms of average and median per-image error, by using different ways of creating the initial shapes. Specifically, we show the performance when using a single initialisation, when using multiple initialisations generated only by shifting the mean

[3]We have also tested a version of the code provided by the authors, which results in similar performance.




(a) LFPW dataset

(b) Helen dataset




(c) IBUG dataset

Figure 4: Cumulative iod-normalised error curve for 49 inner-facial landmarks for different datasets. The red line corresponds to the SDM, the black line corresponds to the cascaded regression using the $L_{2,1}$-normalised error, and the green line corresponds to the proposed method.

| | | | |
|---|---|---|---|
| LFPW | 0.038 (0.035) | 0.042 (0.039) | 0.046 (0.039) |
| Helen | 0.041 (0.038) | 0.049 (0.042) | 0.059 (0.043) |
| IBUG | 0.163 (0.088) | 0.0162 (0.090) | 0.207 (0.130) |

Table 1: Mean (median) of the per-image error when using different initialisation strategies. Left column: the initialisation used in our approach. Centre column: multiple initialiastions constructed by shifting a frontal exemplar. Right column: Only one initial shape (see text for more details on these approaches).

shape fitted to the bounding box, and when using our approach. The latter creates initial shape hypothesis by fitting 4 reference shapes to the bounding box (shown in Fig. 3) and then perturbing them spatially. It is interesting to note here that the computational complexity of the algorithm depends to a large degree on the number of initial shapes considered. The number of evaluations in the fist case is 375 ($9\times9\times4+7\times7+1+1$), 132 in the second ($9\times9+7\times7+1+1$) and 4 in the latter case. Since our implementation did run comfortably within the limits set by the organisers, we decided to maximise performance.

*Performance per iteration.* The graphs shown in Fig. 5 depict the per-iteration cumulative error graph for the test partitions of the LFPW and Helen datasets, and for the IBUG dataset. This includes the error resulting from fitting the mean shape to the detection bounding box. This is the best guess based only on the face bounding box, and provides a good measure of the accuracy and robustness of the face detector. However, in practise we use multiple initialisations. There is convergence after only three iterations. We perform a fourth one, giving a marginal increment (curves of the last two iterations are mostly overlapped in Fig. 5).

Again, these graphs are constructed exactly with the parameters and code submitted to the challenge (except for the correction of the face detection bug). Thus, the parameters were not optimised or tweaked in any way to provide the best performance on these datasets. It would be however possible to use the statistics of the training partition of the LFPW or Helen dataset to obtain

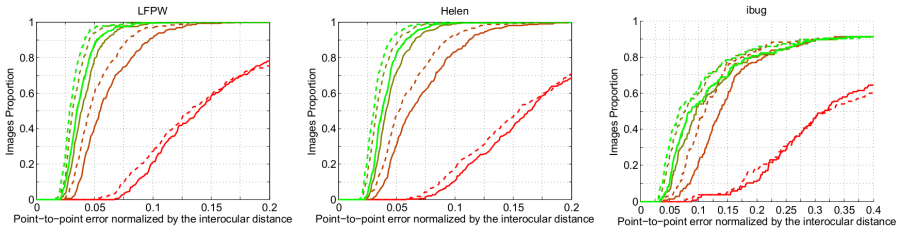



Figure 5: Cumulative iod-normalised error on the LFPW (left), Helen (centre) and IBUG (right) datasets. Dashed lines are for the inner facial landmarks, while continuous lines are the error for all landmarks. The colour code indicates the iteration: red is the face detection, then for each iteration the colour moves towards green. The final detection is depicted in full green.

better parameters for the test set, while increasing the size of the perturbations for the IBUG dataset would likely lead to better performance. We consider however more fair to provide results with exactly the same parameters for all datasets. Please note that, for Fig. 5, the maximum error shown in the graphs is of 0.2 for LFPW and Helen, and of 0.4 for the IBUG dataset.

*300W challenge results.* As previously mentioned, after submission we found a bug on the test function for the face detector. The results shown throughout the paper were obtained with the corrected version. However, the challenge results, shown in what follows, were obtained with a version of the code that included the bug.

## 6. Conclusions

In this article, we have tackled the problem of facial landmarking in the wild by focusing on augmenting the robustness of current methods to non-frontal head poses. While we build on the hugely popular SDM, we have two major contributions to differentiate this work from previous ones. Experimental results confirm that the resulting algorithm is indeed very robust. This results in particularly good performance for the most challenging datasets.

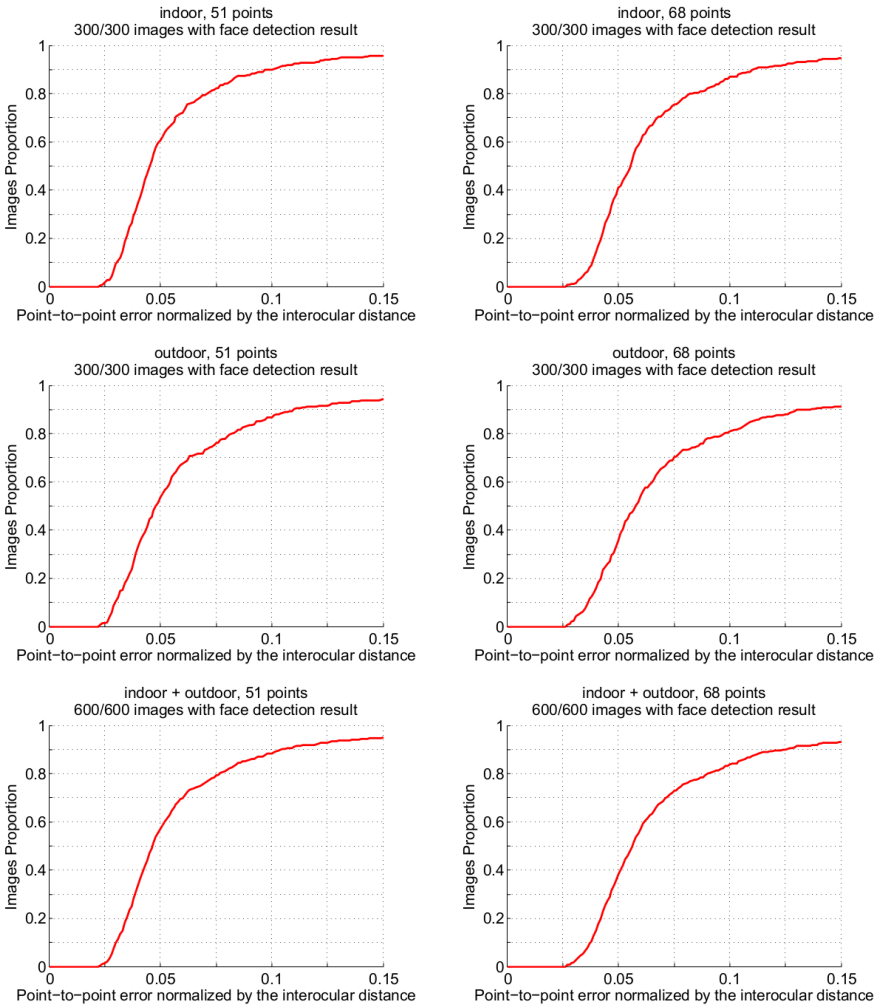



Figure 6: Results on the 300W challenge, divided by inlaying points (51 landmarks, left column) and all (68 landmarks, right column), and between indoor images (upper row), outdoor images (central row) and both combined (lower row)

**Acknowledgements**

The work of Dr. Valstar and Dr. Martinez is funded by European Union Horizon 2020 research and innovation programme under grant agreement No. 645378 ARIA-VALUSPA.

**References**

[1] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, Trans. on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.

[2] I. Matthews, S. Baker, Active appearance models revisited, Int'l Journal of Computer Vision 60 (2) (2004) 135–164.

[3] G. Tzimiropoulos, M. Pantic, Optimization problems for fast AAM fitting in-the-wild, in: Int'l Conf. Computer Vision, 2013.

[4] J. M. Saragih, S. Lucey, J. F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int'l Journal of Computer Vision 91 (2) (2011) 200–215.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, Training models of shape from sets of examples, in: British Machine Vision Conf., 1992.

[6] P. Belhumeur, D. Jacobs, D. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, Trans. on Pattern Analysis and Machine Intelligence 35 (12) (2013) 2930–2940.

[7] A. Asthana, S. Cheng, S. Zafeiriou, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Computer Vision and Pattern Recognition, 2013, pp. 3444–3451.

[8] M. F. Valstar, B. Martinez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, in: Computer Vision and Pattern Recognition, 2010, pp. 2729–2736.

[9] B. Martinez, M. F. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression based facial point detection, Trans. on Pattern Analysis and Machine Intelligence 35 (5) (2013) 1149–1163.

[10] T. F. Cootes, M. C. Ionita, C. Lindner, P. Sauer, Robust and accurate shape model fitting using random forest regression voting, in: European Conf. on Computer Vision, 2012, pp. 278–291.

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

[12] P. Dollár, P. Welinder, P. Perona, Cascaded pose regression, in: Computer Vision and Pattern Recognition, 2010, pp. 1078–1085.

[13] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Computer Vision and Pattern Recognition, 2013.

[14] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 FPS via regressing local binary features, in: Computer Vision and Pattern Recognition, 2014, pp. 1685–1692.

[15] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Computer Vision and Pattern Recognition, 2014.

[16] X. P. Burgos-Artizzu, P. Perona, P. Dollár, Robust face landmark estimation under occlusion, in: Int'l Conf. Computer Vision, 2013, pp. 1513–1520.

[17] J. Xing, Z. Niu, J. Huang, W. Hu, S. Yan, Towards multi-view and partially-occluded face alignment, in: Computer Vision and Pattern Recognition, 2014, pp. 1829–1836.

[18] G. Tzimiropoulos, Project-out cascaded regression with an application to face alignment, in: Computer Vision and Pattern Recognition, 2015, pp. 3659–3667.

[19] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Computer Vision and Pattern Recognition, 2013, pp. 3476–3483.

[20] J. Yan, Z. Lei, D. Yi, S. Li, Learn to combine multiple hypotheses for accurate face alignment, in: Int'l Conf. Computer Vision Workshop, 2013, pp. 392–396.

[21] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: the first facial landmark localization challenge, in: Int'l Conf. Computer Vision Workshop, 2013.

[22] J. Liu, S. Ji, J. Ye, Multi-task feature learning via efficient $L_{2,1}$-norm minimization, in: Conf. on Uncertainty in Artificial Intelligence, 2009, pp. 339–348.

[23] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, Trans. on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645.

[24] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[25] J. Orozco, B. Martinez, M. Pantic, Empirical analysis of cascade deformable models for multi-view face detection, Image and Vision Computing.

[26] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005, pp. 886–893.

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

[28] V. Le, J. Brandt, Z. Lin, L. D. Bourdev, T. S. Huang, Interactive facial feature localization, in: European Conf. on Computer Vision, 2012, pp. 679–692.

[29] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: Computer Vision and Pattern Recognition, 2012, pp. 2879–2886.

[30] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Incremental face alignment in the wild, in: Computer Vision and Pattern Recognition, 2014.

[31] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, A semi-automatic methodology for facial landmark annotation, in: Comp. Vision and Pattern Recog. - Workshop, 2013.