Pattern Recognition Letters
journal homepage: www.elsevier.com

Cascaded Regression with Sparsified Feature Covariance Matrix for Facial Landmark Detection

Enrique Sánchez-Lozano[**], Brais Martinez, Michel F. Valstar

School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, U.K.

ABSTRACT

The recent appearance of regression-based methods that directly infer the whole shape has revolutionised the facial landmarking problem, and they have quickly become the state-of-the-art approach. The most notable exemplar is the Supervised Descent Method. Its main characteristics are the use of the cascaded regression approach, the use of the full appearance as the inference input, and the aforementioned aim to directly predict the full shape. In this article we argue that the key aspects of the algorithm are the use of cascaded regression and the avoidance of the constrained optimisation problem that characterised most of the previous approaches. Instead, we show that, surprisingly, it is possible to achieve comparable or superior performance using only landmark-specific predictors and linearly combining them. We reason that augmenting the input with context (of which using the full appearance is the extreme) can sometimes be harmful. In fact, we found that there is a relation between the data variance and the benefits of adding context to the input. We finally devise a simple greedy procedure that makes use of this fact to obtain superior performance to the SDM, yet maintaining the simplicity of the algorithm. We show extensive results both for intermediate results devised to prove the main aspects of the argumentative line, and to validate the overall performance of two models constructed based on these considerations.

1. Introduction

Existing facial landmark detection approaches are commonly divided into part-based and holistic approaches. Holistic approaches are almost restricted to the Active Appearance Models family (Cootes and Taylor (2001), Matthews and Baker (2004)). They represent the full face appearance, and are typically generative. Facial landmarking results in this case as a by-product of the dense reconstruction of the face appearance. Instead, part-based models are characterised by representing the face as a constellation of patches around the facial landmarks. They are typically discriminative (Saragih et al., 2011), although it is also possible to use part-based generative models (Tzimiropoulos and Pantic, 2014). While generative methods are capable of attaining very precise results when the search is initialised close to the solution (Tzimiropoulos and Pantic, 2013), discriminative methods provide better robustness. In this article we focus on part-based discriminative models, as they are the most widely used. Many of the existing works on part-based facial landmarking can be cast in the Constrained Local Models (CLM) framework[1] introduced by Saragih et al. (2011).

The CLM framework devises landmark detection as the iterative alternation between two step, the response map construction and the response maximisation step. Response maps encode the likelihood of any given image location of being the true landmark location, and a different response map is constructed for each landmark. Many works used classifiers to create these landmarks (e.g. Saragih et al. (2011); Belhumeur et al. (2011); Asthana et al. (2013, 2015)). A probabilistic classifier (e.g., a logistic regressor) can be trained as to distinguish the true landmark location from surrounding locations. At test time, the classifier can be evaluated over a region of interest on a sliding window manner. The response map is then con-

[**]Corresponding author: Tel.: +44-115-823-228; fax: +44-115-951-4254;
e-mail: psxes1@nottingham.ac.uk (Enrique Sánchez-Lozano),
Brais.Martinez@nottingham.ac.uk (Brais Martinez),
Michel.Valstar@nottingham.ac.uk (Michel F. Valstar)
URL: www.cs.nott.ac.uk/~exs (Enrique Sánchez-Lozano)

[1] The term Constrained Local Model was previously introduced by Cristinacce and Cootes (2006) prior to the work by (Saragih et al., 2011). Furthermore, it has become somewhat common to refer to the specific approach proposed in Saragih et al. (2011) as the CLM, while their method was introduced only as a particular instance of the CLM framework (the authors called their method Regularised Landmark Mean Shift. In this article we refer to CLM as the general framework rather than to any specific methodology.

structed using the predicted likelihoods. The response maximisation step consists of finding the valid shape maximising the combined per-landmark responses. Thus, this step is a maximisation constrained by the shape model. The shape fitting step is very challenging, and it contains multiple local minima. Thus, many authors have focused their efforts on improving this step. For example, Saragih et al. (2011) attained real-time reliable fitting by using a Mean Shift-constrained optimisation. However, the Mean Shift optimisation is prone to getting stuck in local maxima, especially for the flexible shape parameters, responsible for expressions. To overcome this, Belhumeur et al. (2011) proposed a variation of RANSAC, so that a very large number of solutions were generated using training set exemplars. Then, the highest-scoring ones were linearly combined into the final solution. Asthana et al. (2013) used instead discriminatively trained regressors to find adequate increments to the shape parameters, and Asthana et al. (2015) proceeded by training a generative model of the response maps and then using it to perform the maximisation.

Recent years have seen however the appearance of works employing regressors instead of classifiers to exploit local appearance (Valstar et al., 2010). It was soon shown that the regressors resulted in improved response maps and hence better global performance (e.g. Cootes et al. (2012); Martinez et al. (2013)). However, a constrained optimisation problem was still necessary in order to enforce shape consistency, consequently hindering performance. Further performance improvement was attained by considering regressors trained to directly infer the full shape increments necessary to move from the current shape estimate to the ground truth. That is to say, instead of using the appearance of a single landmark to predict only over this landmark, they used the full appearance to predict over the full shape, eliminating the need for a subsequent step enforcing shape consistency. This was pioneered by Cao et al. (2012), which also proposed the use of cascaded regression (Dollár et al., 2010) to this end. However, it was the Supervised Descent Method (SDM) (Xiong and De la Torre, 2013) that became the de-facto state of the art. While they maintained the main concepts of Cao et al. (2012), they simplified the method by using Least Squares for regression, and concatenated per-landmark HOG features as their feature representation. This resulted in a very simple algorithm (feature extraction aside, only 4 matrix multiplications are involved!) capable of attaining the best performance to date.

What the key advantages are of the SDM respect to other methods is thus a relevant question. Several factors characterise the algorithm: the cascaded regression, the use of context (i.e., the concatenation of all the local descriptors into a single feature vector), and the direct prediction of the shape. Each one has arguably some merit. The cascaded regression allows for combined robustness and precision, the use of context provides an input with augmented descriptive power, and the direct shape increment prediction removes the need for subsequent complex optimisation steps.

In here we argue that only two of these components, the cascaded regression and the direct estimation of the shape, are enough as to produce similar or even better results to those of the SDM. That is to say, if these two aspects are respected, similar performance can be attained with and without context. We further investigate at which extent the use of context within the input features is necessary, exploring intermediate solutions between landmark-independent predictions and the SDM approach. In order to eliminate context from the regression models, we resort to the sparsification of the feature covariance matrix. We show experiments highlighting the relation between the amount of context used (i.e., the sparseness of the feature covariance matrix), and the variance of the data. Finally, we use this relation to build a variant of the SDM algorithm with decreasingly sparse matrices at each iteration. This algorithm can be very easily implemented given an SDM implementation, has less computational complexity, and achieves superior performance in practise. We use the LFPW, Helen, AFW and IBUG datasets (see Sec. 6 for details) to validate the analysis and to show practical performance of the solution derived from it.

A previous version of this manuscript appeared in Sánchez-Lozano et al. (2013). The work presented in this article differs from it in that we provide a different interpretation and mathematical derivation to justify the matrix sparsification, provide a link between the benefits of sparsification and data variance that was missing in the previous version, we link the success of direct regression-based methods with the avoidance of constrained optimisation, and we have completely revamped the introduction and experimental results section.

2. Cascaded Linear Regression

Let $\mathbf{I}$ be a face image, for which we want to estimate the ground truth shape $\mathbf{s}^{gth}$, consisting of n facial landmarks (thus being a $2n$-dimensional vector). Let $\mathbf{s}$ be an estimation of the location of these points, then $\boldsymbol{\phi}(\mathbf{I}, \mathbf{s}) \in \mathbb{R}^{p\times 1}$ represents the features extracted around the positions defined by $\mathbf{s}$ within image $\mathbf{I}$. The feature vector is constructed by extracting a HOG descriptor at a small patch centred around each landmark, and then concatenating them into a single feature vector. The regression target is defined as $\mathbf{y} = \mathbf{s}^{gth} - \mathbf{s}$. That is to say, $\mathbf{y}$ is the increment necessary to move from the current estimate $\mathbf{s}$ to the ground truth shape $\mathbf{s}^{gth}$. It is then possible to define a linear regressor $\{\mathbf{R}, \mathbf{b}\} \in \{\mathbb{R}^{2n\times p}, \mathbb{R}^{2n\times 1}\}$ tasked with translating image features into shape increments. Specifically, the increment $\mathbf{y}$ is estimated as $\mathbf{R}\boldsymbol{\phi}(\mathbf{I}, \mathbf{s}) + \mathbf{b}$ and the updated shape estimate is computed as $\mathbf{y} + \mathbf{s}$. This linear regressor can be expressed in a more compact form by defining $\tilde{\boldsymbol{\phi}}(\mathbf{I}, \mathbf{s})$ as the result of adding a one to the end of $\boldsymbol{\phi}(\mathbf{I}, \mathbf{s})$. Then, $\tilde{\mathbf{R}}$ is defined as a $\mathbb{R}^{2n\times p+1}$ matrix, so that:

$$\mathbf{R}\boldsymbol{\phi}(\mathbf{I}, \mathbf{s}) + \mathbf{b} = \tilde{\mathbf{R}}\tilde{\boldsymbol{\phi}}(\mathbf{I}, \mathbf{s}) \tag{1}$$

The data variance is in practise too large as to attain a good estimate of the true shape using only one single regressor. This limitation is overcome through the use of the cascaded regression. The idea is to sequentially apply a set of regressors rather than using a single one. At test time, an initial shape estimate $\mathbf{s}^0$ is computed using the face detection bounding box. Then, the cascaded regression produces a sequence of estimates as

$\mathbf{s}^k = \mathbf{s}^{k-1} + \tilde{\mathbf{R}}^k \tilde{\boldsymbol{\phi}}(\mathbf{I}, \mathbf{s}^{k-1})$. If the cascade has N_k iterations, then $\mathbf{s}^{N_k}$ is the estimate of $\mathbf{s}^*$.

The training of the cascade starts with a *data augmentation* strategy (Dollár et al., 2010), which proceeds by generating m different initial shapes for each of the n_{im} training images. These shapes can for example be generated by aligning a reference shape (e.g. the mean shape) to the ground truth by means of a translation and scaling. Then, the aligned reference shape is perturbed in terms of translation and scaling, sampling the perturbation uniformly within a range. This results in a set of initial shapes $\mathbf{s}^0_{i,j}$. The Least Squares regressor k is then computed as:

$$\tilde{\mathbf{R}}^k = \arg\min_{\mathbf{R}} \sum_{i}^{n_{im}} \sum_{j}^{m} ||\mathbf{s}_i^{gth} - \mathbf{s}_{i,j}^{k-1} - \mathbf{R}\, \tilde{\boldsymbol{\phi}}(\mathbf{I}_i, \mathbf{s}_{ij}^{k-1})||_2^2 \qquad (2)$$

and the training shapes for the next iteration are defined by applying the trained regressor to each of the training shapes as:

$$\mathbf{s}_{i,j}^{k} = \mathbf{s}_{i,j}^{k-1} + \tilde{\mathbf{R}} \tilde{\boldsymbol{\phi}}(\mathbf{I}_i, \mathbf{s}_{i,j}^{k-1}) \qquad (3)$$

The minimisation in Eq. 2 is simply a least squares equation, and it has a closed form solution. We first set the notation $\mathbf{X}^k$ as the matrix that results from storing the vectors $\tilde{\boldsymbol{\phi}}(\mathbf{I}_i, \mathbf{s}_{i,j}^{k-1})$ as its columns. Furthermore, we consider that all the features are normalised as to have 0 mean and standard deviation 1 across all the training set, except for the feature corresponding to the bias term. Similarly, $\mathbf{Y}^k$ is defined as the matrix containing $\mathbf{s}_i^{gth} - \mathbf{s}_{i,j}^{k-1}$ on its columns. Then the optimal regressor is defined as:

$$\tilde{\mathbf{R}}^k = \mathbf{Y}^k \mathbf{X}^{k^T} \left(\mathbf{X}^k \mathbf{X}^{k^T} \right)^{-1} \qquad (4)$$

It is interesting to note that, despite the joint form of the prediction function, each of the outputs is estimated independently of one another. That is to say, if we were to define $2n$ regressors taking the same input as in Eq. 4, but where the target would be a 1-dimensional, the output would be the same. Thus, SDM does not enforce shape consistency. Instead, the output is (approximately) valid due to the use of the same input for each of the $2n$ regressors[2].

3. Context vs. no context

We observed that the SDM formulation is actually equivalent to training a different regressor to predict each of the $2n$ dimensions of the output. The use of the full appearance as the input can be interpreted as the use of context. That is to say, the prediction for a specific landmark is not computed only with landmark-specific information, but rather with information regarding any other landmark. In this section we argue that the key aspect is the use of the same input for each of the output dimension-specific regressors rather than the use of context in the input. To this end, we will define an algorithm that uses no context (prediction is based only on landmark-specific information). We will show that this algorithm attains equal or even superior performance to the SDM despite not using context.

Let us note $\mathbf{C} = \mathbf{X}\mathbf{X}^T$ as the covariance feature matrix involved in Eq. 4 (we are ignoring the index of the cascade iteration for simplicity of notation). We further note $\mathbf{C}^{i,j}$ as the covariance matrix between the features resulting from landmark i and those resulting from landmark j. $\mathbf{X}^i$ is defined too as a block of $\mathbf{X}$ containing the features associated with landmark i. Let us devise an algorithm parallel to the SDM, but where no context is used to perform prediction. More specifically, let us obtain a prediction of the full shape in exactly the same way, but now only using the appearance of a single landmark as the input. This same landmark-specific prediction can be obtained for each landmark, resulting in n predictions. Finally, we combine all of the n predictions into a single one by computing a weighted mean of the landmark-specific predictions. This can be specified in mathematical terms by first defining the per-landmark predictions *of the full shape* as:

$$\hat{\mathbf{y}}_{*,l} = \tilde{\mathbf{R}}_l \mathbf{x}_*^l = \mathbf{Y}\mathbf{X}^{l^T} \left(\mathbf{C}^{l,l} \right)^{-1} \mathbf{x}_*^l \qquad (5)$$

where $l = \{1, \dots, n\}$ indexes the landmarks, we use the asterisk for variables defined for the test image, and $\hat{\mathbf{y}}_{*,l}$ is the prediction of the full shape generated using the appearance of landmark l. Then the test shape estimate for the next stage of the cascade is defined as follows:

$$\mathbf{s}_* + \hat{\mathbf{y}}_* = \mathbf{s}_* + \sum_{l=1}^{n} w_l \hat{\mathbf{y}}_{*,l} \qquad (6)$$

This process is very similar to previous landmarking methods (e.g., Valstar et al. (2010)). However, it scraps the alternation between per-landmark predictions and shape-level constrains by performing a prediction over the full face shape at once (although the multiple predictions have to be combined). Now let us represent this into a more compact equation as:

$$\hat{\mathbf{y}}_* = \sum_{l=1}^{n} w_l \mathbf{Y}\mathbf{X}_l{}^T \left(\mathbf{C}^{l,l} \right)^{-1} \mathbf{x}_*^l = \mathbf{Y}\mathbf{X}_l^T \tilde{\mathbf{C}}^{-1} \mathbf{x}_* \qquad (7)$$

where:

$$\tilde{\mathbf{C}} = \begin{pmatrix} w_1^{-1}\mathbf{C}^{1,1} & 0 & \cdots & 0 \\ 0 & w_2^{-1}\mathbf{C}^{2,2} & 0 & \cdots \\ \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & w_{n+1}^{-1}\mathbf{C}^{n+1,n+1} \end{pmatrix}, \qquad (8)$$

It is interesting to now note that the prediction for the standard SDM regression takes a very similar form to Eq. 7. The only difference is that $\mathbf{C}$ is now substituted by a much sparser matrix $\tilde{\mathbf{C}}$, where all relations between features associated to different landmarks are set to 0.

In here we interpret this relation as follows: the SDM makes full use of the context in the data representation, and this is reflected in the dense feature covariance matrix associated to its formulation. Instead, the use of independent regressors is equivalent to the use of a block-diagonal (i.e. very sparsified)

[2] While it has been argued that the output is consistent due to the fact that shapes lie in a linear subspace, this is actually not precise. Shapes are assumed to lie on a linear subspace once rigid transformations are removed. However, it can be argued that the shape subspace is approximately locally linear.

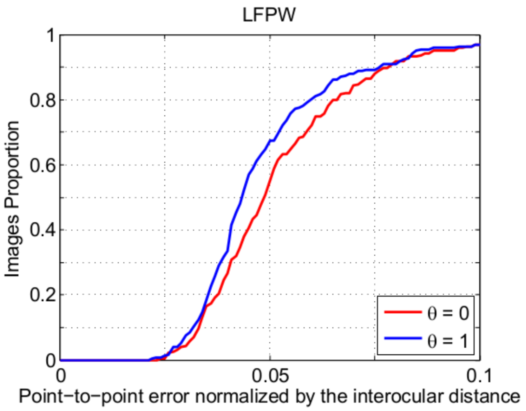


Fig. 1: Performance (in terms of the iod-normaliased error) for the SDM (red) and the averaging of landmark-independent predictions (blue) on the test partition of the LFPW dataset.

matrix. However, there are intermediate levels of sparsification of **C**, each one corresponding to a different level of context. In the following we define the general case, of which the SDM and the landmark-independent approaches are special cases.

We performed an experiment to see the impact of the use of context on the quality of the prediction. The performance of both algorithms was measured on the LFPW test partition. The prediction error was computed using the Inter-Ocular Distance (iod)- normalised measure (see Sec. 6 for details). The resulting cumulative error distributions for both the SDM and the landmark-independent methods are shown in Fig. 1. It is possible to see that, surprisingly, using the sparse matrix $\tilde{\mathbf{C}}$ actually results in slightly better performance than using the full covariance matrix, especially in terms of robustness. That is to say, the use of the full context does not help!.

4. Sparsifying the covariance matrix

In this section we explore levels of context intermediate between it being used in full, and it being totally discarded. To this end, let us express the full feature covariance matrix as the block-wise matrix:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}^{1,1} & \dots & \mathbf{C}^{1,n+1} \\ \vdots & \ddots & \vdots \\ \mathbf{C}^{n+1,1} & \dots & \mathbf{C}^{n+1,n+1} \end{pmatrix} \tag{9}$$

We can now define the Pearson correlation coefficient between patches based on the blocks of **C** as:

$$\gamma_{i,j} = \frac{||\mathbf{C}^{i,j}||_F^2}{||\mathbf{C}^{i,i}||_F||\mathbf{C}^{j,j}||_F} \tag{10}$$

The value $\gamma_{i,j} \in [0,1]$ defines how correlated the features corresponding to landmarks i and j are throughout the dataset. If the aim is to drop the less useful context from the features, this can be done by eliminating the less correlated blocks within **C**. We can now sparsify **C** by suppressing every $\mathbf{C}^{i,j} \quad s.t. \quad \gamma_{i,j} < \theta$, where $\theta \in [0,1]$. We note the resulting sparsified feature covariance matrix as $\mathbf{C}_\theta$. Please note that the resulting matrix is still a valid covariance matrix since $\gamma_{i,j} = \gamma_{j,i}$. An example of the feature covariance matrix, and the matrix of coefficients $\gamma_{i,j}$, are illustrated in Fig. 2. If $\theta = 0$, then $\mathbf{C}_{\theta=0} = \mathbf{C}$, and the method reduces to the standard SDM. Instead, if $\theta = 1$, then $\mathbf{C}_{\theta=1} = \tilde{\mathbf{C}}$, and the resulting method is the landmark-independent regression method.

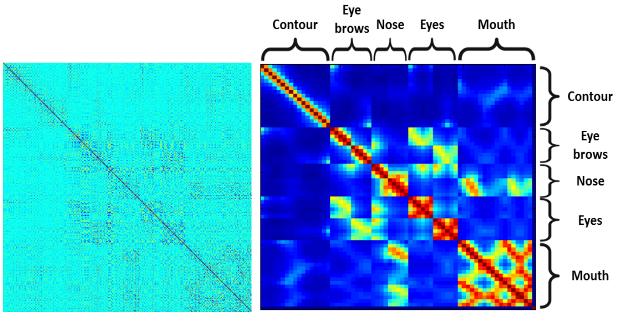


Fig. 2: Examples of feature covariance matrix (left) and patch-based correlation coefficients (right, red indicates higher correlation, blue lower; better seen in colour).

As the matrix becomes sparser, different disjoint components appear. Each of these components produce a separate prediction of the full shape, similarly as indicated in 5, except that now there are less disjoint components than landmarks. Still, each component produces a shape prediction, and they need to be linearly combined. The prediction for each of the components can be computed as a closed form solution (see Eq. 4). The mixing values $\{w_i\}_{i=1:N_c}$, where N_c indicates the number of components, are found by keeping a portion of the training data for validation purposes (a full crossvalidation could be similarly used). After optimal parameters are found, the training of the least squares regressor is conducted with the full amount of data. The disjoint components in the covariance matrix can be found through a simple clustering algorithm using the correlations $\gamma_{i,j}$ as input. The output are index sets $\mathcal{S}$. The coefficients in **C** corresponding to each of these clusters for an independent component, and thus can produce its own prediction of the full shape independently of the rest of the elements of the covariance matrix.

Algorithm 1 Find disjoint components on **C**

```
Require: $\{\gamma_{i,j}\}_{i,j=1...n}$.
Compute $\mathbf{Z} = \{z_{i,j} = 1 \,|\, \gamma_{i,j} \geq \theta\}_{i,j=1,...,n}$
Compute $\mathcal{S}_i = \{\, j \,|\, z_{i,j} = 1\}_{i=1,...,n}$
for i = 1...n do
    for j = i+1...n do
        if $\mathcal{S}_i \cap \mathcal{S}_j > \emptyset$ then $\mathcal{S}_i \leftarrow \mathcal{S}_i \cup \mathcal{S}_j$, $\mathcal{S}_j \leftarrow \emptyset$
        end if
    end for
end for
return $\{\mathcal{S}_i \neq \emptyset\}$
```

5. Sparsity and Input Data Variance

The optimal amount of context used on each iteration of the cascade might vary. In here we hypothesise that the use of context becomes harmful when the data added does not correlate well with the current patterns. The role of context is to disambiguate, i.e., to clarify which examples are really similar and which are not. However, adding patterns that are loosely correlated to the existing input can introduce a confusing signal instead of helping disambiguate. That is to say, adding features can either disambiguate if both signals collaborate to identify similar examples, or can corrupt the input if both signals disagree with each other.

In order to study the correlation between data variance and the level of sparsity, we designed the following experiment. We trained 5 different models (see Sec. 6 for details on the experimental setup). The first model was trained using decreasing sparsity values for each level of the cascade. Specifically, since there are 5 iterations of the cascade, we select thresholds $\theta = 1$ to $\theta = 0$ with decrements of 0.25 at each iteration. The second model was trained inverting the order of the thresholds (i.e., values go from $\theta = 0$ to $\theta = 1$ this time). Finally we trained three other models where the thresholds were kept set to $\theta = 0$, $\theta = 0.75$ and $\theta = 1$ respectively throughout the cascade. The results for the LFPW and Helen datasets are shown in Fig. 3. It is again clear from this analysis that the best performing methodology is to increase the level of context (decrease the sparsification parameter) for each iteration of the cascaded regression. The first model results in better performance than any other model. Instead, the second model (augmenting the thresholds for each cascaded regression iteration) results in the opposite, and produces the worst performance of all models.

6. Experiments

This section contains the experimental results showing the practical gain attained by sparsifying the feature covariance matrix. We compare three models, the first two trained with feature matrix sparsification. We use however two different criteria to define the thresholds. The first one is constructed using a greedy parameter search. That is to say, the optimal parameters for iteration 1 are computed irrespectively of the parameters in successive stages. Then this parameter is fixed, and we proceed to find the optimal parameter for the next level of the cascade. This same procedure is followed for all the levels. The parameters for a given level are found using a grid search. The automatically found sparsification thresholds for each iteration of the cascade are 0.65, 0.8, 0.75, 0.1 and 0. The trend of decreasing the sparsification parameter for later levels of the cascade is clear. Our second model uses the simple heuristic of Sec. 5. Specifically, we assign θ with decreasing values, from 1 to 0 at a stride of 0.25.

We further compare performance with the SDM. This would be equivalent to using a constant $\theta = 0$. We however also compute PCA on the input data, keeping 98% of the energy. While this improves performance, PCA cannot be easily applied to the sparsification approach, as PCA entangles the features. Thus, it would not be possible to find a correspondence between features and landmarks. It is however only fair to compare our method against a version of the SDM including this step.

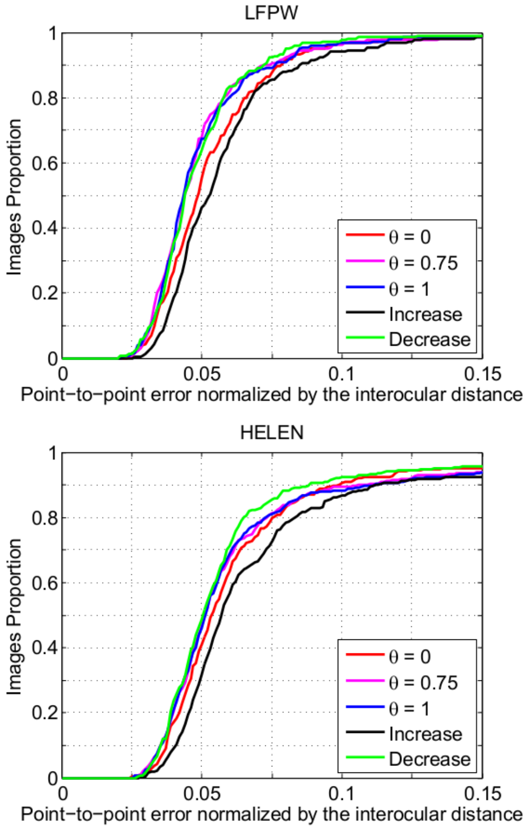


Fig. 3: Cumulative error distribution for the LFPW (up) and Helen (bottom) datasets. Results were obtained with 5 different heuristics for defining the sparsity threshold at different levels of the regression cascade. The black curve shows increments of 0.25 from 0 to 1, the green curve shows the inverse pattern (decreasing from 1 to 0). The other three curves show performance using constant thresholds.

Datasets:. We use the training partition of the LFPW Belhumeur et al. (2011) for the training of our models. The tests are carried out on the testing partitions of the LFPW and the Helen datasets, as well as on the AFW Zhu and Ramanan (2012) and IBUG datasets (see the 300W challenge[3]). By doing so, we have four datasets of increasing difficulty. While the LFPW and the Helen datasets are of similar complexity, the test on the LFPW dataset is dataset-dependent. Instead, the test on the Helen dataset is dataset-independent. The AFW dataset is more challenging than both the LFPW and Helen datasets, while the IBUG dataset is extremely challenging. We used the re-annotations provided for the 300W challenge Sagonas et al. (2013), although we use 66 landmarks instead of the 68 annotated.

Initial shape:. The initialisation is based on the bounding box automatically found by a face detector trained using the Deformable Parts Model (there was however minimal intervention in the cases the right face was not the highest scored one). For training, data augmentation is used to produce 10 initial shapes per image.

[3] `http://ibug.doc.ic.ac.uk/resources/300-W_IMAVIS/`

Error measure:. For every test dataset, we construct a cumulative error distribution. Every point on the y axis shows the percentage of test examples where the detection yielded an error bellow the x axis value. The error is measured as the mean point-to-point Euclidean distance, normalised by the interocular distance as defined on the 300W challenge (i.e., using the outer eye corners). Some other works have used the distance between the centres of the eyes (computed as the mean of each eye corners), or the average face bounding box size. This results in significantly different error scales, and this should be beared in mind when interpreting these type of graphs.

Results:. By training the sparsity parameters on the LFPW, we have tuned our algorithm for a dataset of similar difficulty. This is shown in the the performance gain on the test partition of the LFPW dataset. The model with heuristically-defined sparsity parameters yields good although slightly inferior performance. Similar relative performances can be observed in the Helen dataset. Instead, the AFW dataset shows slightly worse performance when using the greedily-found parameters. For all these cases, the SDM algorithm performs worse than any of the two models proposed. The IBUG dataset however is much more challenging than LFPW. Thus, the levels of sparsity defined on that dataset are no longer ideal. As a result, the SDM performs similarly to this model. Instead, the heuristics with which the second model was trained are not data-dependent, and this model still comes atop in terms of performance. It is important to note that while all the graphs shown in this article have a maximum error of 0.15, the graph corresponding to the IBUG dataset has a maximum error of 0.4. This is due to its very challenging nature, resulting in the cumulative error function stabilising at higher error values. These results highlight two costributions of this article, the usefulness of sparsifying the feature covariance matrix, and the association between the need for context and the variability of the data.

We further provide qualitative results in Fig. 5. They serve to illustrate the nature of the datasets employed, and the practical meaning of the error values. We provide 3 images per dataset, where the alignment was obtained by the model using heuristically-defined thresholds. The last image for each dataset reflects an alignment failure. It is interesting to note that the LFPW dataset contains few non-frontal head poses and thus most of the errors happen on these cases.

7. Conclusions

In this article we examine some of the reasons behind the recent success of the SDM, specifically focusing on its use of context. We show that a full use of context is not ideal, explore different intermediate levels by sparsifying the feature covariance matrix, and show the relation between context and the data variance. Specifically, the major conclusions of this article are: 1) the use of context is not always beneficial, and similar or even superior performance can be attained without the use of context; 2) We show instead that defining the target of inference as the full shape is a key algorithmic aspect; 3) this implies that the view of facial landmarking as a constrained optimisation problem, which has been widely accepted until very recently, is actually inadequate in practise; 4) we reason about the relation of the training data variance and the need for context within the inputs. We also show experimental evidence that strongly suggests context is beneficial in the presence of higher data variances. 5) We train and evaluate a model trained in a greedy manner as to pick the right amount of context for each iteration. We show that this simple trick improves the performance of the SDM significantly. We will also release a binary executable, on the author's website, to facilitate the testing of the proposed method (upon acceptance).

Acknowledgments

This work was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378. The work of Enrique Sánchez-Lozano has been partially funded by the fellowships program of the Fundación Barrié.

References

Asthana, A., Cheng, S., Zafeiriou, S., Pantic, M., 2013. Robust discriminative response map fitting with constrained local models, in: Computer Vision and Pattern Recognition.

Asthana, A., Zafeiriou, S., Tzimiropoulos, G., Cheng, S., Pantic, M., 2015. From pixels to response maps: Discriminative image filtering for face alignment in the wild. Trans. on Pattern Analysis and Machine Intelligence .

Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N., 2011. Localizing parts of faces using a consensus of exemplars, in: Computer Vision and Pattern Recognition.

Cao, X., Wei, Y., Wen, F., Sun, J., 2012. Face alignment by explicit shape regression, in: Computer Vision and Pattern Recognition.

Cootes, T., Taylor, C., 2001. Active appearance models. Trans. on Pattern Analysis and Machine Intelligence 23, 680–689.

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

Cristinacce, D., Cootes, T.F., 2006. Feature detection and tracking with constrained local models, in: British Machine Vision Conf., pp. 929–938.

Dollár, P., Welinder, P., Perona, P., 2010. Cascaded pose regression, in: Computer Vision and Pattern Recognition.

Martinez, B., Valstar, M., Binefa, X., Pantic, M., 2013. Local evidence aggregation for regression-based facial point detection. Trans. on Pattern Analysis and Machine Intelligence 35, 1149–1163.

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

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

Sánchez-Lozano, E., Argones-Rúa, E., Alba-Castro, J., 2013. Blockwise linear regression for face alignment, in: British Machine Vision Conf.

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

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

Tzimiropoulos, G., Pantic, M., 2014. Gauss-newton deformable part models for face alignment in-the-wild, in: Computer Vision and Pattern Recognition.

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

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

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

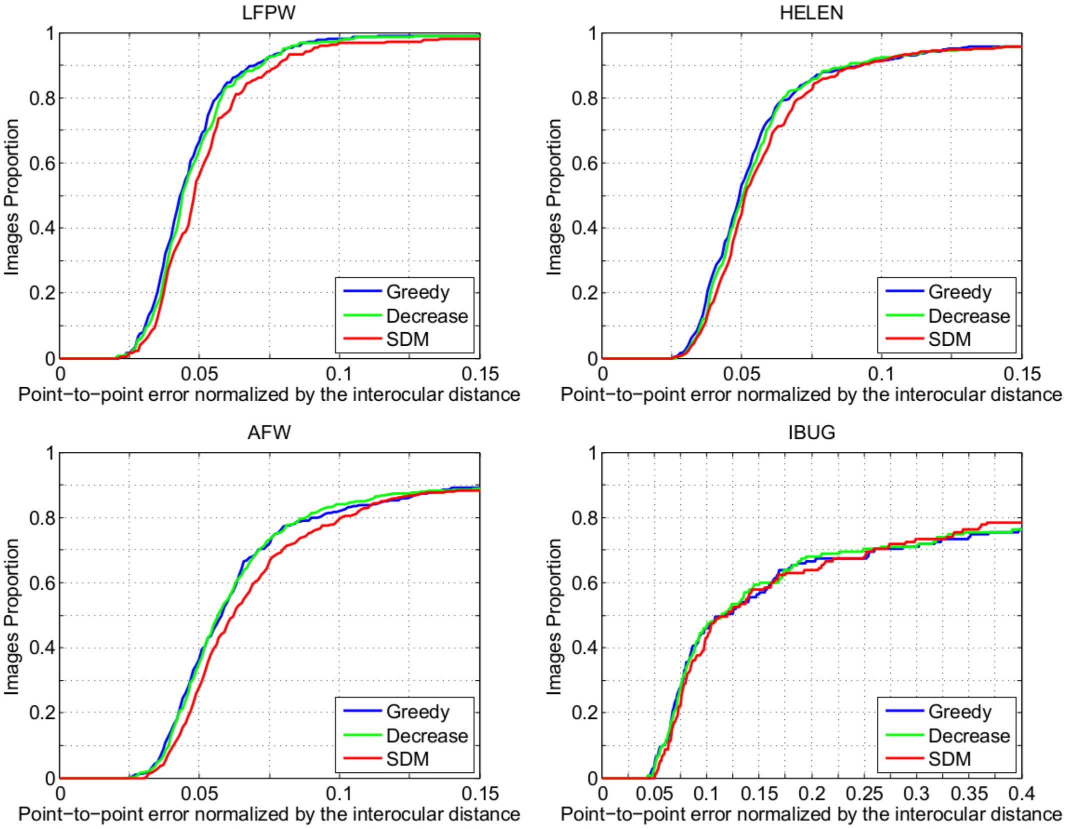
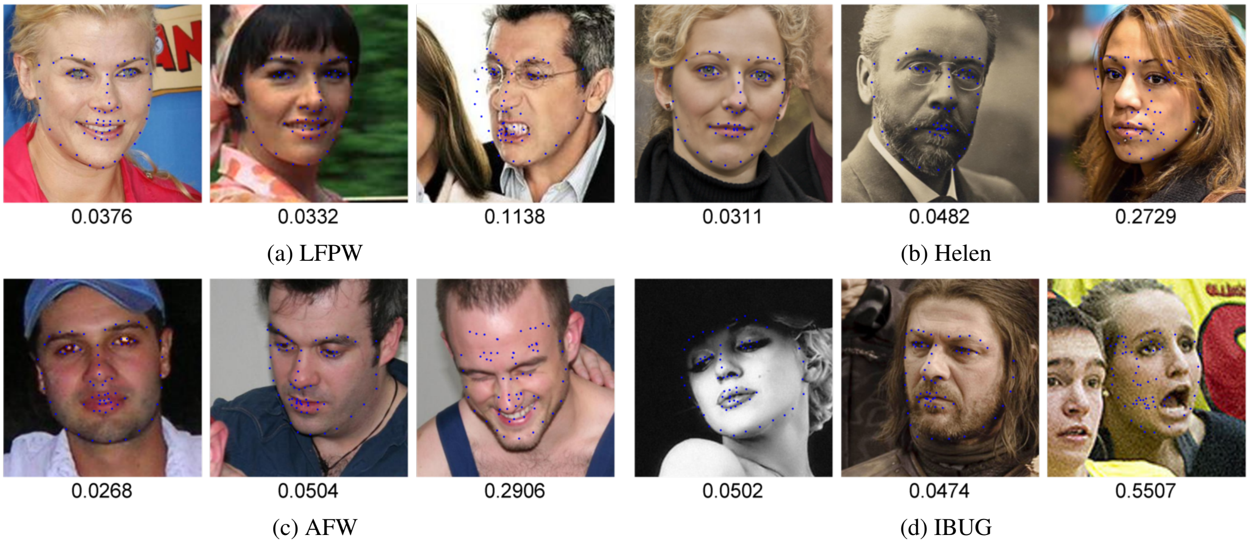


Fig. 4: Performance per method for each dataset. Red indicates the SDM performing PCA on the feature space, blue has the greedy search PO method, green corresponds to the model trained with heuristically-defined descending values for θ.



(a) LFPW

(b) Helen

(c) AFW

(d) IBUG

Fig. 5: Qualitative results on all datasets used. The last image for each dataset shows a fitting failure. The IOD error for each image are shown below each corresponding image.