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

This paper explores the use of context on regression-based methods for facial landmarking. Regression based methods have revolutionised facial landmarking solutions. In particular those that implicitly infer the whole shape of a structured object have quickly become the state-of-the-art. The most notable exemplar is the Supervised Descent Method (SDM). 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 responsible for the success of SDM are the use of cascaded regression and the avoidance of the constrained optimisation problem that characterised most of the previous approaches.We show that, surprisingly, it is possible to achieve comparable or superior performance using only landmark-specific predictors, which are linearly combined. We reason that augmenting the input with too much context (of which using the full appearance is the extreme case) can be harmful. In fact, we experimentally 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, while maintaining the simplicity of the algorithm. We show extensive results both for intermediate stages 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

Structured object detection is an active research area in Computer Vision, where the aim is to describe the shape of an object by locating its parts. Facial landmark detection is a prime example of this, and it is a key step in many applications such as face recognition or facial expression recognition, where the alignment step based on the location of the parts is crucial to achieve a good performance.

Existing facial landmark detection approaches are commonly divided into part-based and holistic approaches. Holistic approaches are mostly 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, each centred 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 steps, response map construction and response maximisation. 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

**Corresponding author: Tel.: +44-115-823-2287; fax: +44-115-951-4254;
e-mail: Michel.Valstar@nottingham.ac.uk (Michel F. Valstar)
URL: www.cs.nott.ac.uk/~mfv (Michel F. Valstar)

[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. In this article we refer to CLM as the general framework rather than to any specific methodology.

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 to distinguish the true landmark location from surrounding locations. At test time, the classifier can be evaluated over a region of interest in a sliding window manner. The response map is then constructed 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 converge at 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. The highest-scoring exemplars were linearly combined into the final solution. Asthana et al. (2013) instead used 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 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 the location of this landmark, the full appearance is used to predict the entire shape, eliminating the need for a subsequent step enforcing shape consistency. This was pioneered by Cao et al. (2014), who 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. (2014), 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 capable of attaining the best performance to date (only 4 matrix multiplications are involved, not counting feature extraction!).

Is thus an important line of investigation to analyse what the key advantages are of the SDM with respect to other methods. Several factors characterise the algorithm: the cascaded regression, the implicit 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 can be argued to have 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.

We argue that using only two of these components, to wit the cascaded regression and the direct estimation of the shape, is sufficient 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 to 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 variability of the data in terms of factors such as the head pose, image quality, facial expressions or identity. 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 more complete 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, and we link the success of direct regression-based methods with the avoidance of constrained optimisation.

The contributions of this work can thus be summarised as:

- We analyse which are the key methodological aspects behind the performance success of the SDM.
- We show that, surprisingly, we achieve superior performance to the standard SDM when encoding no context within the input features.
- We show that there is an inverse correlation between the benefits of using context and the variance of the input data.
- Based on these observations, we devise a simple yet effective extension of the SDM, where each regressor uses an optimal amount of context within the input features. The resulting method is shown to outperform SDM

2. Cascaded Linear Regression

Let $\mathbf{I}$ be a face image, for which we want to estimate the ground truth shape $\mathbf{s}^g$, consisting of n facial landmarks (thus being a $2n$-dimensional vector). Let $\mathbf{s}$ be an estimation of the location of these points, then $\boldsymbol{\phi}(\mathbf{I}, \mathbf{s}) \in \mathbb{R}^{p\times 1}$, with p the dimension of the feature space, 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 features of all patches into a single feature vector. The regression target is defined as $\mathbf{y} = \mathbf{s}^g - \mathbf{s}$. That is to say, $\mathbf{y}$ is the increment necessary to move from the current estimate $\mathbf{s}$ to the ground truth shape $\mathbf{s}^g$. 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 to attain an accurate prediction of the true shape using only a single prediction made by one single regressor. In the SDM, 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 iterations, then $\mathbf{s}^N$ 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 transformation. 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:

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

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

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

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 $\tilde{\boldsymbol{\phi}}(\mathbf{I}_i, \mathbf{s}^{k-1}_{i,j})$ as its columns. Furthermore, we consider that all the features are normalised 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 $\mathbf{s}^g_i - \mathbf{s}^{k-1}_{i,j}$ on its columns. Then the optimal regressor is defined as:

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

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 1-dimensional, the output would be the same. Thus, SDM does not enforce shape consistency. Instead, the output shape is (approximately) valid due to the use of the same input for each of the $2n$ regressors

3. Context vs. no context

We observe 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 all landmarks. In this section we argue that the key aspect is that the same input $\boldsymbol{\phi}(\mathbf{I}, \mathbf{s})$ is used for each of the n output dimension-specific regressors, rather than that $\boldsymbol{\phi}(\mathbf{I}, \mathbf{s})$ actually encodes context. 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 compared to the SDM despite not using any context at all.

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). $\mathbf{X}^i$ is defined too as a block of $\mathbf{X}$ containing the features associated with landmark i. We further note $\mathbf{C}^{i,j}$ as the sub-matrix of $\mathbf{C}$ between the features resulting from landmark i and those resulting from landmark j. 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:

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

where $l = \{1, \ldots, n\}$ 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:

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

This process is very similar to previous landmarking methods (e.g., Valstar et al. (2010)). However, it does not alternate between per-landmark predictions and shape-level constrains, instead performing a prediction over the full face shape at once, through the combination of multiple predictions. Now let us represent this into a more compact equation as:

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

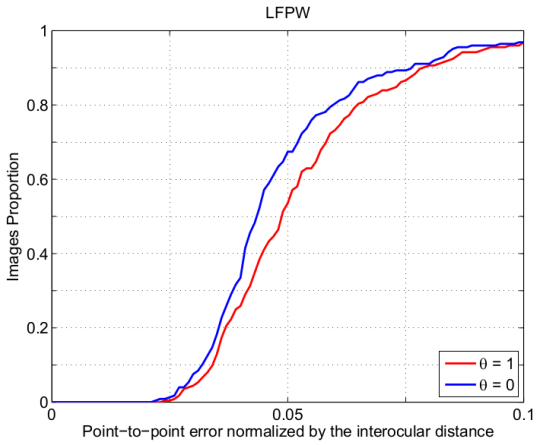

Fig. 1: Performance (in terms of the IOD-normalised error - see Sec. 6 for details) for the SDM (red) and the averaging of landmark-independent predictions (blue) on the test partition of the LFPW dataset. The y value of the graph indicates the percentage of test images with an error equal or lower to the x value.

where:

$$\tilde{\mathbf{C}} = \begin{pmatrix} w_1^{-1}\mathbf{C}^{1,1} & 0 & \cdots & 0 \\ 0 & w_2^{-1}\mathbf{C}^{2,2} & 0 & \cdots \\ \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & w_n^{-1}\mathbf{C}^{n,n} \end{pmatrix}, \tag{8}$$

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 landmark-independent regressors, i.e., regressors that use only appearance information from one landmark to predict, is equivalent to the use of a block-diagonal (i.e. very sparse) matrix. However, there are intermediate levels of sparseness of $\mathbf{C}$, 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. The level of context is defined by the context parameter $\theta \in [0, 1]$. $\theta = 0$ corresponds to not using any context, while $\theta = 1$ corresponds to the full use of context (i.e. SDM). 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!

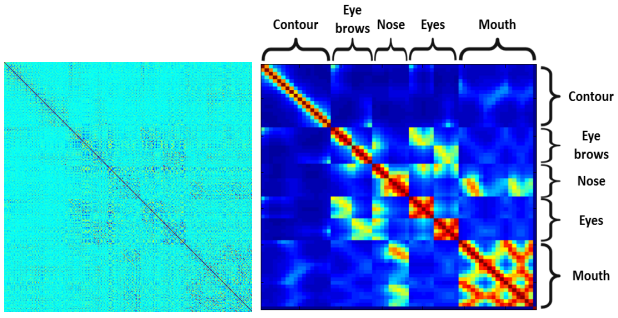

Fig. 2: Examples of feature covariance matrix (left) and patch-based correlation coefficients, derived from Eq. 10 (right, red indicates higher correlation, blue lower; better seen in colour).

4. Sparsifying the covariance matrix

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

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

We can now define the Pearson correlation coefficient between patches based on the blocks of $\mathbf{C}$ 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 remove the least useful context from the features, this can be done by eliminating the least correlated blocks within $\mathbf{C}$. We can now sparsify $\mathbf{C}$ by suppressing every $\mathbf{C}^{i,j}$ for which $\gamma_{i,j} < 1 - \theta$, where $\theta \in [0, 1]$ is the level of context. We denote the resulting sparsified feature covariance matrix as $\mathbf{C}_\theta$. Please note that the resulting matrix can still be treated as a 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 $\theta = 1$, then $\mathbf{C}_{\theta=1} = \mathbf{C}$, and the method reduces to the standard SDM. Instead, if $\theta = 0$, then $\mathbf{C}_{\theta=0} = \tilde{\mathbf{C}}$, and the resulting method is the landmark-independent regression method.

As the matrix becomes sparser, different disjoint components appear [2]. 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

[2] Further mathematical derivation, as well as how to find these disjoint components, can be found in Sánchez-Lozano et al. (2013).

used). After optimal parameters are found, training of the least squares regressor is conducted with the full amount of data.

It is interesting to note in here that several works have tackled the problem of feature selection following a sparse-inducing criterion Zhang et al. (2011); Xu et al. (2014). Our work differs from these in that we generate sparsity on the covariance matrix rather than on the original feature space. Furthermore, we do not resort to the commonly-used L_1 norm regulariser to enforce the sparsity, but rather eliminate entries in a block-wise manner within the covariance matrix to increase the level of sparsity of the matrix (i.e., reduce the number of non-zero entries).

5. Sparsity and Input Data Variability

In order to study the correlation between input data variability and the level of sparsity, we designed the following experiment. We trained 5 different models (see Sec. 6 for details on the experimental set-up). 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 $\theta = 1$, $\theta = 0.25$ and $\theta = 0$ respectively throughout the cascade. The results for the LFPW and Helen datasets are shown in Fig. 3. These results gives experimental evidence that increasing the level of context for each iteration of the cascaded regression (i.e., increase θ) improves the performance. The second model results in better performance than any other model. Instead, the first model (decreasing the thresholds for each cascaded regression iteration) has the opposite effect, and produces the worst performance of all models.

The optimal amount of context used on each iteration of the cascade might vary. We speculate 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 help to 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.

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. When this parameter is determined, 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.

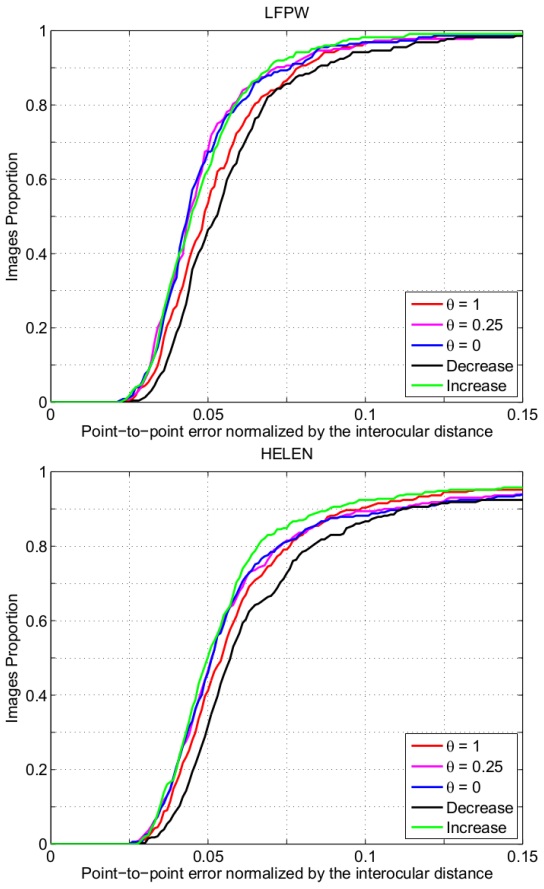

Fig. 3: Cumulative error distribution for the LFPW (up) and Helen (bottom) datasets (see Sec. 6 for details on the error measure). The y value of the graph indicates the percentage of test images with an error equal or lower to the x value. The sparsity threshold were defined with 5 different heuristics. Black: decrements of 0.25 from 1 to 0; Green: increments of 0.25 from 0 to 1; The other three curves use constant thresholds.

The sequence of automatically found sparsification thresholds for each iteration of the cascade are 0.35, 0.2, 0.25, 0.9 and 1. The trend of an increasing sparsification parameter, and thus increased use of context, for later levels of the cascade, is clear. Our second model uses the simple heuristic of Sec. 5. Specifically, we assign θ with increasing values, ranging from 0 to 1 at a stride of 0.25.

We further compare performance with the SDM. This would be equivalent to using a constant $\theta = 1$. 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 projects the input shape space onto a low dimensional space, in which dimensions have no physical meaning. 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 dimensionality reduction step, since by using PCA SDM achieves better results.

Datasets:. We use the training partition of 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 (Le et al., 2012) datasets, as well as on the AFW (Zhu and Ra-

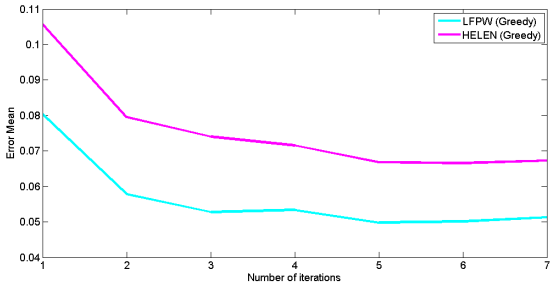

Fig. 4: Average mean error (y axis) vs. number of iterations of the cascade (x axis) on the LFPW and Helen datasets.

manan, 2012) and IBUG (Sagonas et al., 2013b) datasets (see the 300W challenge[3]). By doing so, we have four datasets of increasing difficulty. While LFPW and Helen are of similar complexity, the test on LFPW is dataset-dependent. In contrast, the test on Helen 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., 2013b), 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 For training, data augmentation is used to produce 10 initial shapes per image.

Error measure:. The error is measured as the mean point-to-point Euclidean distance, normalised by the interocular distance as defined on the 300W challenge (Sagonas et al., 2013a), i.e., defined as the distance between the outer eye corner landmarks. For every test dataset, we further construct a cumulative error distribution. Every point on the y axis shows the percentage of test examples where the detection yielded an error below the x axis value. 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.

Number of cascade iterations:. In here we study how many iterations of the cascade should be computed, i.e., we empirically fix the parameter N. To this end, we have tested the performance of the Greedy-search model using different number of iterations. In this model, the ideal threshold is found automatically using a CV strategy. The results for the LFPW and Helen datasets are shown in Fig. 4. From this experiment, we conclude that 5 iterations provide the ideal error. This is one more iteration than the one reported for the SDM (Xiong and De la Torre, 2013).

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

	Greedy	Increase	SDM
LFPW	0.049 (0.043)	0.050 (0.045)	0.054 (0.048)
Helen	0.066 (0.049)	0.067 (0.050)	0.069 (0.052)
AFW	0.107 (0.058)	0.106 (0.057)	0.108 (0.062)
IBUG	0.229 (0.117)	0.226 (0.112)	0.229 (0.124)

Table 1: Mean (median) of the per-image error comparing different approaches.

Results:. Quantitative performance results for our methods and the SDM are summarised in Fig. 5, where the cumulative error distribution is provided, and in Table 1, where the mean and median errors of the different methods on the test datasets are given. The median is given in this case to avoid the over-proportioned impact of large image errors (failed fittings) on the mean error.

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 contributions 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. 6. They serve to illustrate the nature of the datasets employed, and the practical meaning of the error values. 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 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: IEEE Conf. on 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: IEEE Conf. on Computer Vision and Pattern Recognition.

Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. Int'l Journal of Computer Vision 107.

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: IEEE Conf. on Computer Vision and Pattern Recognition.

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

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., 2013a. 300 faces in-the-wild challenge: the first facial landmark localization challenge, in: Int'l Conf. Computer Vision Workshop.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013b. A semi-automatic methodology for facial landmark annotation, in: IEEE Conf. on Computer Vision and Pattern Recognition - 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: IEEE Conf. on 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: IEEE Conf. on 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: IEEE Conf. on Computer Vision and Pattern Recognition.

Xu, J., Yang, G., Yin, Y., Man, H., He, H., 2014. Sparse-representation-based classification with structure-preserving dimension reduction. Cognitive Computation 6, 608–621.

Zhang, D., Zhu, P., Hu, Q., Zhang, D., 2011. A linear subspace learning approach via sparse coding, in: Int'l Conf. Computer Vision, pp. 755–761.

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

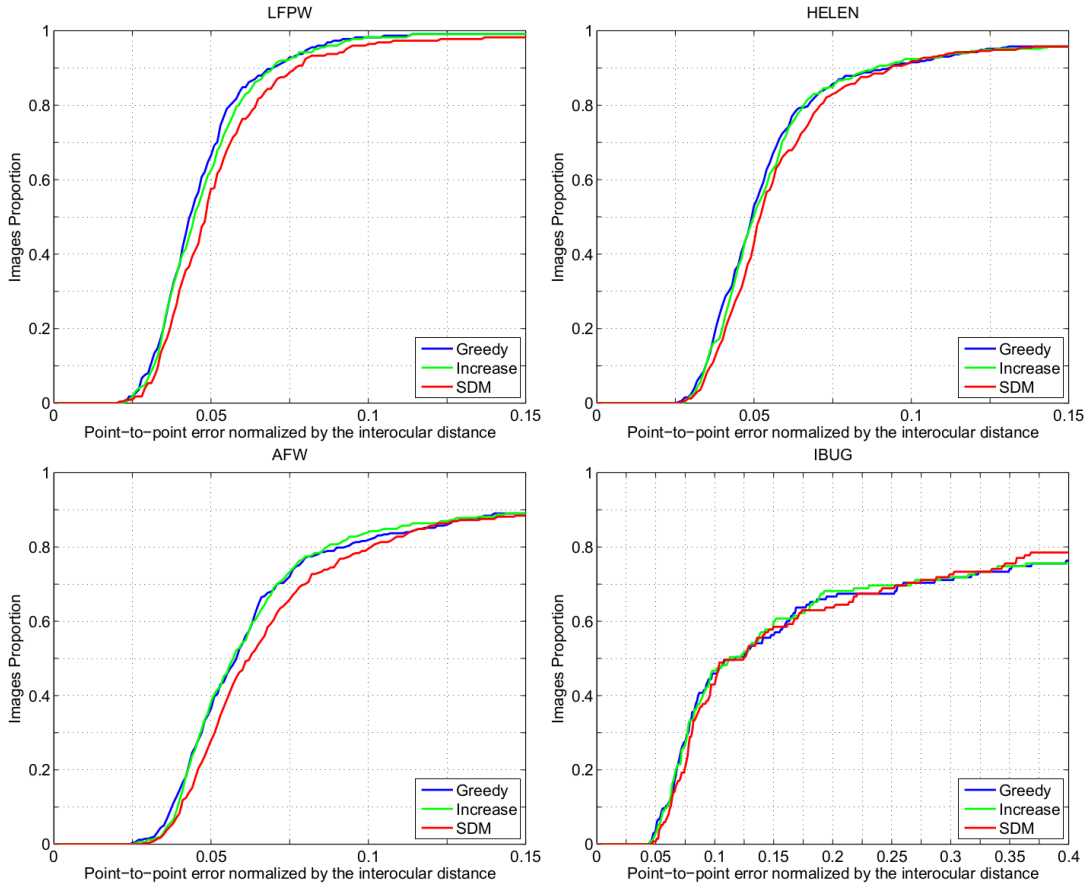

Fig. 5: Performance per method for each dataset. Red indicates the SDM performing PCA on the feature space, blue has the greedy search PO method, green corresponds to the model trained with heuristically-defined increasing values for θ.

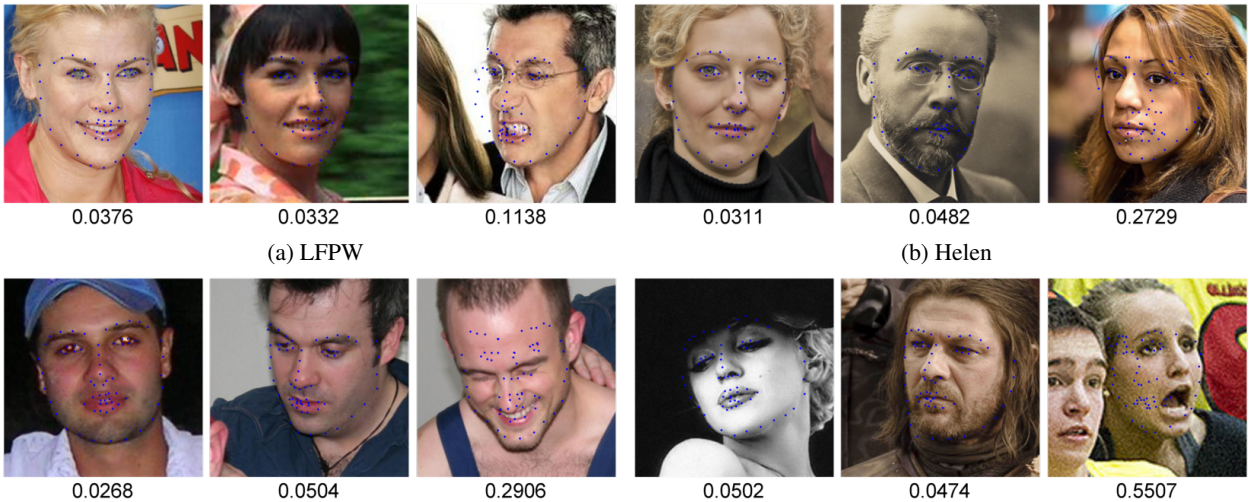

Fig. 6: Qualitative results on all datasets used. The last image for each dataset shows a fitting failure. The IOD error for each image are shown below each corresponding image.