Continuous Supervised Descent Method for Facial Landmark Localisation

Marc Oliu[1,4], Ciprian Corneanu[2,4], László A. Jeni[3], Jeffrey F. Cohn[3,5], Takeo Kanade[3] and Sergio Escalera[2,4]

[1]Universitat Oberta de Catalunya, 156 Rambla del Poblenou, Barcelona, Spain
[2]Universitat de Barcelona, 585 Gran Via de les Corts Catalanes, Barcelona, Spain
[3]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[4]Computer Vision Center, O Building, UAB Campus, Bellaterra, Spain
[5]Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA

Abstract. Recent methods for facial landmark location perform well on close-to-frontal faces but have problems in generalising to large head rotations. In order to address this issue we propose a second order linear regression method that is both compact and robust against strong rotations. We provide a closed form solution, making the method fast to train. We test the method's performance on two challenging datasets. The first has been intensely used by the community. The second has been specially generated from a well known 3D face dataset. It is considerably more challenging, including a high diversity of rotations and more samples than any other existing public dataset. The proposed method is compared against state-of-the-art approaches, including RCPR, CG-PRT, LBF, CFSS, and GSDM. Results upon both datasets show that the proposed method offers state-of-the-art performance on near frontal view data, improves state-of-the-art methods on more challenging head rotation problems and keeps a compact model size.

1 Introduction

Facial landmark location consists of detecting a set of particular points on the face. Usually these points have semantic meaning, their location being in highly distinctive places around the eyes, mouth or nose. A set of such points is useful for expressing both the rigid and non-rigid deformations of the face geometry. Because facial geometry changes with identity, facial expression and head pose, it is an important step in many automatic facial analysis tasks such as face recognition, face expression recognition, face synthesis and age or gender estimation [1].

A common approach for locating landmarks on the face is to model the relation between the face appearance and its geometry. If we consider $\mathbf{X}^*$ to be the ground truth geometry, and $\Phi(\mathbf{I}, \mathbf{X})$ a representation function of a geometry $\mathbf{X}$ on an image $\mathbf{I}$, then starting from an initial estimation $\mathbf{X}^0$ landmark location can be formulated as an optimisation problem of the form:

$$\arg\min_{\Delta\mathbf{X}} f(\mathbf{X}^0 + \Delta\mathbf{X}) = ||\Phi(\mathbf{I}, \mathbf{X}^0 + \Delta\mathbf{X}) - \Phi(\mathbf{I}, \mathbf{X}^*)||_2^2 \quad (1)$$

Because Φ is a highly non-linear function, f is non-convex and has many local minima, the problem becomes severe in the case of large variations of the texture which is normally the case with rotations of the head and strong non-rigid deformations. Additionally, successfully solving the optimisation problem is highly dependent on the initialisation.

Historically, Active appearance models (AAM) [2] are one of the most used methods for 2D face registration. They are an extension of active shape models (ASM) [3] which encode both geometry and intensity information. More recently, even though single step landmark location methods have been proposed [4, 5], the most common approach is to model the relationship between texture and geometry with a cascade of regression functions [6–12]. Features are extracted from the current estimated geometry and passed to the learnt mapping in order to update the geometry. This process is repeated iteratively for each step of the cascade, applying a specific mapping to each. If we denote by $\mathbf{R}^i$ the regression function at the ith step of the cascade, by $\mathbf{\Phi}^i = \Phi(\mathbf{I}, \mathbf{X}^i)$ the corresponding representation and by $\mathbf{b}^i$ a constant bias, then at every step of the cascade, the geometry $\mathbf{X}$ will be updated in the following way:

$$\mathbf{X}^{i+1} = \mathbf{X}^i + \mathbf{R}^i\mathbf{\Phi}^i + \mathbf{b}^i \quad (2)$$

While most cascaded regression methods share this approach, considerable variation can be found in representation, regression functions and initialisation strategies. The simplest way to initialise the geometry is by starting with the mean [9, 11, 8]. For faces, this works well in close-to-frontal scenarios but proves inefficient when large pose variation occurs. A common solution is to try a set of random initialisations and consider the median of the predictions as the final solution [13, 6]. Unfortunately, this considerably increases the computational cost. An alternative approach is to apply the initial part of the cascade and continue only if the variance of the regressed shapes is low, which is a strong predictor of convergence towards the global minimum [7]. If this is not the case then a different set of initial shapes is generated. Even so, all these methods are dependent on the initialisation and prove low generalisation to large head pose rotation. A coarse-to-fine searching approach was recently proposed to deal with the initialisation dependency problem [14]. A regression function is learnt from a set of shapes generated according to a probabilistic distribution on the shape space. A dominant set approach is used to eliminate outliers between the regressed shapes in an unsupervised manner. From the filtered subset the centre of a smaller region of the original space is computed and the process repeated until convergence. While it prevents locality of the solutions it improves robustness to large pose variation.

The work of Dollar et al. which proved influential in the field of facial landmark localisation, uses intensities of sparse sets of pixels at predefined locations to represent texture in a shape indexed fashion for learning a fixed linear se-

quence of weak regressors [13]. In this way, representation's output depends on both the image data and the current estimate of the geometry. Some of the methods propose to jointly learn the representation and the regression function [6, 7, 11, 8]. In this sense, several shape indexed locations are randomly generated and then selected based on a certain optimisation criteria. Alternatively, local binary features are learnt for each landmark independently [8]. During test, very fast landmark localisation is obtained. In a recent method [15], *Difference of Gaussians* (DoG) features are selectively extracted from locations arranged in a pattern inspired by the human visual system [16]. Learnt trees at early stages tend to select indexed DoG features computed from distant sampling points while trees at later stages tend to use nearby sampling points. Finally, a very common problem of most of the proposed methods, the lack of sensitivity to occlusions is tackled in the work of Burgos et al. [7]. They propose a method that reduces exposure to outliers by detecting occlusions explicitly and using robust shape-indexed features. It incorporates occlusion directly during learning to improve shape estimation.

A distinct group of methods use predefined handcrafted representations while learning the regression function from the data. For example, to overcome the large computational time required by the regression of many generated shapes at each stage more simple descriptors are used in the initial stages when coarse localisation is performed. More complicated representations are used on final stages when fine localisation takes place [14]. A particularly important set of methods that use fixed representations are the ones derived from the *Supervised Descent Method* (SDM) [9]. SDM uses simplified SIFT features and linear regressors. As is the case of previous methods, SDM works well for near frontal faces but fails on strong rotations. To overcome this problem, *Global Supervised Descent Method* (GSDM) [10] introduced an approach which uses a sub-space defined by a set of directions of maximum variance of the training data to partition the original feature space. Each partition shares a similar descent direction for the training instances falling within it. A linear regressor is learnt for each partition. However, GSDM suffers from two main problems. Both the number of training instances and model size increase exponentially with the number of sub-space dimensions.

In order to perform landmark localisation under strong rotations while keeping a fast and compact model, this work proposes a continuous formulation of GSDM. Instead of using the sub-space to partition the feature space as GSDM does, it is used to describe a space of linear regressors. This is equivalent to proposing a regressor which estimates the second derivative of the gradient, instead of the first as a standard linear regressor would (e.g. in SDM). While this formulation may not be as expressive as GSDM, the amount of memory and training instances required increases linearly with the number of dimensions of the sub-space. Also, the proposed formulation defines a specific linear regressor for each instance.

In summary, our list of contributions is as follows:

- we present a method that improves state-of-the-art results on strongly rotated faces
- the trained models are small, the amount of memory and training instances required increase only linearly with the number of dimensions of the subspaces
- the method is fast to train due to its closed form solution
- we have synthesised largest 2D face dataset to date, with a challenging face rotation distribution

The rest of the paper is organised as follows: in Section 2 we formulate the proposed method, in Section 3 we present the experimental analysis and finally, in Section 4, we conclude the paper.

Notations. Vectors ($\mathbf{a}$) and matrices ($\mathbf{A}$) are denoted by bold letters. An $\mathbf{u} \in \mathbb{R}^d$ vector's Euclidean norm is $\|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^{d} u_i^2}$. $\mathbf{B} = [\mathbf{A}_1; \ldots; \mathbf{A}_K] \in \mathbb{R}^{(d_1+\ldots+d_K)\times N}$ denotes the concatenation of matrices $\mathbf{A}_k \in \mathbb{R}^{d_k \times N}$.

2 Continuous Supervised Descent Method

2.1 Second order regressor

The original SDM method [9] is an exemplar-based method which learns a series of linear regressors approximating the data to the global optima in a cascaded manner. Lets consider $\mathbf{X}^i \in \mathbb{R}^{n\times m}$ the m targets for each of n samples at a given cascade step i, $\Delta\mathbf{\Phi}^i \in \mathbb{R}^{n\times(k+1)} = \mathbf{\Phi}^i - \overline{\mathbf{\Phi}^i}$ the difference of the feature vectors of length k from the mean, with a column vector of ones added in order to account for the bias, and $\mathbf{R}^i \in \mathbb{R}^{(k+1)\times m}$ the linear regressor for each of the m parameters. Then the update formula for SDM can be expressed as follows:

$$\mathbf{X}^{i+1} = \mathbf{X}^i + (\mathbf{\Phi}^i - \overline{\mathbf{\Phi}^i})\mathbf{R}^i = \mathbf{X}^i + \Delta\mathbf{\Phi}^i\mathbf{R}^i \tag{3}$$

This can be seen as learning a linear approximation of the first-order partial derivatives for each parameter. These correspond to $\partial\Delta\mathbf{X}^{i+1}/\partial\Delta\mathbf{\Phi}^i_j = \Delta\mathbf{\Phi}^i_j\mathbf{R}^i_j$, with $\mathbf{R}^i$ being the Jacobian matrix, $\Delta\mathbf{\Phi}^i_j$ the jth column of $\Delta\mathbf{\Phi}^i$ and $\mathbf{R}^i_j$ the jth row of $\mathbf{R}^i$. To make this approximation, the slope is considered homogeneous for any point of the feature space. This assumption does not hold for most problems, where the gradient direction suffers from large variations on different locations of that space. On Global SDM [10] these variations are handled by partitioning the space into different regions and learning a linear regressor for each one. This approach can approximate with high accuracy the gradient variations at different regions of the space, but has the problem of doubling the amount of learnt regressors and required training data each time the space is divided.

Here we introduce a continuous formulation, where a set of bases are learnt for the regressors, effectively learning a linear approximation of the second derivative. To do so, first a set of main modes of variation are learnt from either $\Delta\mathbf{X}^*$ or $\Delta\mathbf{\Phi}^i$ using Principal Component Analysis (PCA):

$$\widetilde{\Delta\mathbf{\Phi}^i} = \left[\Delta\mathbf{\Phi}^i\mathbf{P}_{1:l}, \mathbf{1}_n\right], \tag{4}$$

Where l represents the number of bases to learn and $\mathbf{P}_{1:l}$ is the projection matrix. $\mathbf{1}_n \in \mathbb{R}^{n\times 1}$ denotes an all-ones vector. Given that the total number of learnable parameters for one of m targets equals $p = (k+1)(l+1)$, learning the second derivative for all parameters ($l = k$) would drastically increase the problem dimensionality. Estimating the second derivative on the l main variation modes is a more treatable problem. Given one of the targets $\Delta\mathbf{X}_j^i \in \mathbb{R}^{n\times 1}$, its associated second order regressor is expressed as the solution to the following minimisation problem:

$$\underset{\mathbf{R}_j^i}{\arg\min} ||(\Delta\mathbf{\Phi}^i \circ (\widetilde{\Delta\mathbf{\Phi}^i}\mathbf{R}_j^i))\mathbf{1}_{(k+1)} - \Delta\mathbf{X}_j^i||_2^2 \tag{5}$$

Here, $\mathbf{R}_j^i \in \mathbb{R}^{(l+1)\times(k+1)}$ is the set of l bases (and baseline or bias regressor) describing the regressor for the jth target at the ith cascade step, and $\circ$ denotes the Hadamard product. Note that, according to Equation 7, this formulation learns a linear approximation to the second order partial derivatives $\partial^2\Delta\mathbf{X}_j^{i+1}/(\partial\Delta\mathbf{\Phi}_p^i\ \partial\Delta\widetilde{\mathbf{\Phi}}_q^i) = \Delta\mathbf{\Phi}_p^i\Delta\widetilde{\mathbf{\Phi}}_q^i(\mathbf{R}_j^i)_{pq}$. Thus $\mathbf{R}_j^i$ corresponds to a compact version of the Hessian matrix for target j at cascade step i, having the dimensionality of the feature space reduced before applying the second derivative. Equation 5 can be seen as a compact formulation defining a quadratic regressor for each target, which is known to be a linear problem, having a closed form solution. This minimisation problem can be expressed in a least squares form, providing a closed form solution, as follows:

$$\underset{\mathbf{R}_j^i}{\arg\min} ||(\widetilde{\Delta\mathbf{\Phi}^i} \odot \Delta\mathbf{\Phi}^i)vec(\mathbf{R}_j^{i\intercal}) - \Delta\mathbf{X}_j^i||_2^2 \tag{6}$$

Here $\odot$ denotes the Khatri–Rao product, considering each instance (row) on $\Delta\mathbf{\Phi}^i$ and $\widetilde{\Delta\mathbf{\Phi}^i}$ as a partition of the matrix, and $vec(\mathbf{R}_j^{i\intercal}) \in \mathbb{R}^{(kl+2)\times 1}$ is the vectorisation of the regressor bases. Thus, while the second derivative estimate is used for a subset of principal components, the regressor remains linear. This allows us to rapidly and directly find the optimal regression weights given the training instances. Note that this formulation could be extended to estimate higher order derivatives by applying the Khatri–Rao product multiple times. At test time, the parameters are updated with the following equation:

$$\mathbf{X}_j^{i+1} = \mathbf{X}_j^i + (\Delta\mathbf{\Phi}^i \circ (\widetilde{\Delta\mathbf{\Phi}^i}\mathbf{R}_j^i))\mathbf{1}_{(k+1)} \tag{7}$$

This formula estimates the regressor weights and bias for the current value of the principal components $\widetilde{\mathbf{\Phi}^i}$, and applies it to the features. This is more memory-efficient than performing the Khatri-Rao product of $\Delta\mathbf{\Phi}^i$ and $\Delta\widetilde{\mathbf{\Phi}}^i$ and then performing a linear regression. The bias for the regressor bases is the baseline regressor for an instance with the mean value for the l principal components (PCs) of the feature vector. Each of the l regression bases in $\mathbf{R}^i$ corresponds to the second derivative estimate wrt. a given PC. Note that when $l = 0$ the model

is a standard linear regressor. Thus, SDM can be seen as a special case of our method where the second derivative is not taken into account for any PC.

The proposed approach estimates a standard linear regressor for each instance given the coordinates of the features sub-space $\Delta\widetilde{\mathbf{\Phi}}^i$. Global SDM assigns the same one to all instances falling into a given region of the partitioned sub-space. Another advantage of this approach is that the number of parameters learnt p at each cascade step increases linearly with the number of bases ($p = (k+1)(l+1)$). With Global SDM it increases quadratically ($p = (k+1)min(1, l^2)$). These two factors make the proposed approach both more compact in terms of memory and more accurate, as shown in Section 3.3. Because the regression space is continuous, the weights of the linear regressor are adapted to each instance, providing more flexibility to the model. During training, this also implies that for the proposed approach all the training data is available for each base of the sub-space, helping to reduce over-fitting. GDSM distributes the data between quadrants, logarithmically reducing the available training data for each quadrant with the number of sub-space bases.

2.2 Implementation details

As discussed in Section 2.1, the second derivative of the feature space is calculated over the l principal components. For this work, similarly to [9], a simplified SIFT descriptor is extracted from each landmark estimate. The descriptor has a fixed 32×32 window around the landmark, rotated according to the in-plane rotation of the current geometry relative to the mean facial shape. PCA is then applied in order to reduce its dimensionality. Thus, the feature vector for an instance j at the cascade step i is defined as $\mathbf{\Phi}^i_j = sift(\mathbf{I}_j, \mathbf{X}^i_j)^\intercal \mathbf{P}^i_{1:k}$, the k principal components of the extracted SIFT descriptors. This implicitly provides the l parameters for the regressor bases, being $\widetilde{\mathbf{\Phi}^i_j} = (\mathbf{\Phi}^i_j)_{1:l}$. The targets $\Delta\mathbf{X}^i$ are rotated in the same way as the descriptor windows in order to maintain a coherent update direction.

The feature vector length k and number of regression bases l may depend on the problem and are free parameters of the model. Still, there are two considerations to take into account. In a cascaded regression approach, the first steps of the cascade broadly approximate the face pose and general shape, while later steps tend to fine-tune the location of each landmark, working more locally. This implies that at the fist steps a smaller amount of the total descriptors variance may be enough. Conversely, a higher amount of regression bases would increase the adaptability to the descriptors main modes of variation, which are expected to be caused by pose/illumination variations. The feature vector length k is defined as a fixed percentage of the original SIFT features variance. While it may be possible to adjust the number of bases l at each cascade step (for instance with forward selection), in this work a global value is chosen for all cascade steps.

The initial shape at the first cascade step is the mean shape. It is calculated from the training instances ground truth shapes using Generalised Procrustes Analysis.

3 Experiments

This section is dedicated to the description and discussion of the experiments conducted to validate the proposed method. We begin in Section 3.1 by describing the two datasets we used, 300W a dataset intensely used by the community and BU4DFE-S, a dataset we have specially synthesised from BU4DFE, a 3D face dataset. In Section 3.2 we present the experimental setup and the methods used for comparison [1]. In Section 3.3 we discuss the results.

The objective of these experiments are two-fold. First we want to show that the proposed method achieves state-of-the-art results on close-to-frontal faces. For this purpose we use 300W, a well known public dataset which is the de-facto standard benchmarking dataset for facial landmark localisation. We then want to show that the method outperforms other methods when applied to heavily rotated faces. For this purpose we show results on the BU4DFE-S, a dataset specially synthesised for this purpose. The reader is referred to Table 1 for the overall results on the two considered datasets and to Figure3 for a comparative study of the robustness to rotation 3. Detection examples are presented in Figure 4. The code for the experiments is made publicly available.

3.1 Data

In order to test the proposed method we used 300W, a well known facial expression dataset. We also designed a new dataset, which we called BU4DFE-S, consisting of 2D faces synthesised from BU4DFE, a public 3D dynamic facial expression dataset.

300W. The 300 Faces In-the-Wild (300W) [17] database is a compilation of six re-annotated datasets (68 landmarks). Following the same approach as in [8] [15], four of the six datasets are used: AFW [18], LFPW [19], Helen [20] and iBUG [17]. The test data for LFPW and Helen, along with iBug, are used as test. The rest of data is used for training. This provides a total of 3148 and 689 train and test instances. The data is captured outside the lab and it has balanced ethnic and gender distribution. While challenging and diverse, it does not contain far-from-frontal faces and its number of samples is rather low.

BU4DFE-S: While annotated face datasets have become more challenging and diverse in recent years, they still provide a low number of training instances with limited variation in rotation. In order to compare the robustness of the proposed method with state-of-the-art facial landmark localisation methods, we have created BU4DFE-S, a new large 2D dataset synthesised from the publicly available BU4DFE. BU4DFE is a high resolution dynamic 3D facial dataset [21]. 101 subjects of ages between 18 to 45 years old are captured while showing facial expressions in a controlled environment. The 3D facial expressions are captured at 25 frames per second. Each sequence begins with the neutral expression, proceeds to target emotion and then back to neutral. For creating the BU4DFE-S we sample 5 frames from each captured sequence. The sampled 3D frames

[1] Code and data generation script available at https://github.com/moliusimon/csdm

are equally distributed along the sequence, portraying varying intensities of the same expression during onset, apex and offset.

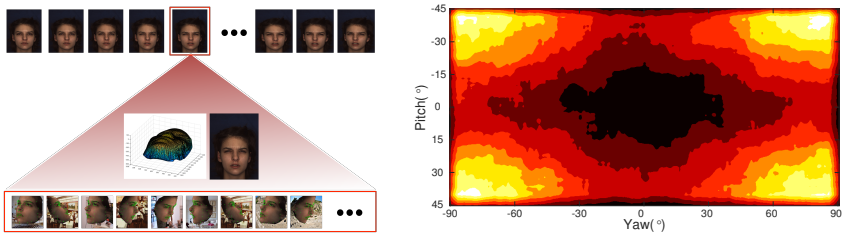
We use the extracted 3D samples, to build 25 2D projections by rotating the 3D model in pitch and yaw. The projected images are generated as follows. The BU4DFE provides the 3D point cloud of the face and an RGB image. Additionally for each of the 3D points the mapping is provided to the corresponding position on the RGB image, making it possible to map the 3D geometry to the colour texture. We first homogeneously down-sample the 3D points set by 20 and build a triangle mesh from the remaining points. The down-sampling factor was heuristically found as a trade-off between the computational cost for generating the projected images and their quality. We consider an isometric projection to associate texture patches to the mesh triangles. The mesh is then rotated with the desired angle and the triangles are again projected to the 2D plane. The new face is built by affine piece-wise warping of the initial texture patches to the newly projections of the rotated triangles by taking into account self occlusions. Inpainting is used to fill warping holes or artifacts. Finally, the images are resized to a standard size of 200×200 pixels and a background is painted on the remaining regions.

We have used the test partition of the Places Dataset, a scene recognition dataset, to build the backgrounds [22]. It contains 41000 images of size 256×256 pixels. From every image we crop two 200×200 regions, one on the top-let corner and the other on the bottom-down corner. The former is flipped. We use these images to place a different background behind each of the generated faces. In Figure 1(a) we provide a summarised depiction of the process. The rotation angles follow an inverse normal distribution for angles between $\pm 90^\circ$ in yaw and $\pm 45^\circ$ in pitch as shown in Figure 1(b). In this way we obtain more highly rotated faces in all directions than close-to-frontal faces. The generated data contains a total of 75000 rotated images of 100 persons. Each person appears in 750 samples with 6 different facial expressions at 5 different intensities rotated 25 times. As the BU4DFE, the subjects are from different ethnicities and follow a balanced gender distribution. The generated dataset has more instances and rotation variation than any other existing public 2D dataset. We show some examples in Figure 2.

Besides containing a larger number of samples (approximately 24 times more than 300W), BU4DFE-S has two more important characteristics. First, for each of the samples the pose is known which is not the case with most of the other 2D face datasets. There exist datasets containing captured faces under different angles in the lab, but the angle distribution is extremely skewed [23]. Another advantage of the BU4DFE-S is that we have total control over the pose distribution of the synthesised data. This makes possible benchmarking the robustness to pose rotation against state-of-the-art methods as shown in Figure 3.

3.2 Experimental settings

For the proposed method the parameter space is larger than for SDM, specially at the first cascade steps. In order to avoid over-fitting, the training data is aug-



(a) Data synthesis for BU4DFE-S. (b) Pose rotation distribution for BU4DFE-S.

Fig. 1: BU4DFE-S contains 2D rotated faces synthesised from BU4DFE, a 3D dynamic facial expression dataset. In (a) we show how from a original sequence we sample a limited number of equally spaced frames. For each of these frames we use the provided 3D mesh and the texture to generate 25 rotated projections. The rotation angle distribution is shown in (b). We favour far-from-frontal faces with respect to close-to-frontal ones in order to make the data as challenging as possible.

mented. For both the 300W and BU4DFE-S datasets the images and geometries are mirrored, doubling the number of training instances. In the case of 300W, which consists of only 3148 training images, the dataset is further augmented by providing 25 different initial geometries. These are generated by applying a random rotation between $[-\pi/4,\ \pi/4]$, a displacement between $[-5\%,\ 5\%]$ for both width and height, and a scaling factor between $[0.9,\ 1.1]$ to the mean shape.

Regarding the number of bases l and the captured feature space variance, the values have been manually chosen for each dataset. For 300W, 2 bases and 95% of variance are used, while for BU4DFE-S, 5 bases and 85% of variance are used. It is necessary to use fewer bases in 300W in order to avoid over-fitting, since the number of training instances is smaller.

We compare the proposed method with the most important facial landmark localisation methods in recent years. This is done using the Normalised Mean Euclidean Error (NMEE) metric, a standard error metric in the literature [9, 7]. It corresponds to the mean euclidean distance between the detected and ground truth landmarks, normalised by the inter-ocular distance. In the case of BU4DFE-S, where large head rotations are present, the 3D inter-ocular distance is used instead. Otherwise for yaw angles close to 90° the inter-ocular distance would tend to zero, giving more weight to errors on heavily rotated faces. For comparing results we considered the most important state-of-the-art methods [6], [7], [9], [10], [11], [8], [20], [14]. RCPR is able to deal with occlusions by including occlusion ground-truth of the landmarks in the learning process. As none of the considered datasets has annotated occlusions we discarded this feature during training. For the ERT [11] and LBF [8], we compare with already published results for the 300W. For a fair comparison we compare the results for the SDM and the GSDM after training with the same number of steps as the proposed method. It is important to note that GSDM is a method oriented to tracking

Fig. 2: BU4DFE-S dataset samples. Annotated face landmarks are shown in green.

the facial geometry, but can be easily applied to the static case by modifying the definition of the subspace used to partition the feature space. Instead of using two principal components from $\Delta\mathbf{X}^i$ and one from $\Delta\mathbf{\Phi}^i$, all principal components are taken from $\Delta\mathbf{\Phi}^i$. For the proposed approach, a 2-dimensional subspace is used in the case of 300W, and a 5-dimensional one for BU4DFE-S. Finally, two recent methods, CFSS [14] and CGPRT [15], have been considered. In their paper, the authors of CGPRT publish two results, with different number of training steps. The result we have obtained was with the larger number of steps and by initialising with the mean shape. CFSS does a constraint search of the shape in a coarse-to-fine manner in subsequent finer shape subspaces. Even though a parallel training on the CPU was attempted, we found training to be very slow, which made impossible obtaining results for BU4DFE-S with the available hardware resources.

3.3 Results discussion

For the 300W dataset, the trained model has been fit to the test data both using mean shape initialisation and with 25 random initialisations sampled using the same criteria used during training (see Section 3.2). The results of both approaches are shown in Table 1. Without multiple test initialisations, the method has a NMEE lower than those achieved by ESR, RCPR, SDM and GSDM, also surpassing ERT when using multiple initialisations. Yet LBF, CGPRT and CFSS still have lower errors. Thus, the proposed approach surpasses, or is close, to most state-of-the-art methods in the near frontal view conditions of the 300-w dataset.

On BU4DFE-S the proposed method outperforms all considered state of the art approaches. Because it is a dataset with large head pose rotations in both

[2] For the BU4DFE-S we compute the interocular distance in 3D.

	ESR [6]	RCPR [7]	SDM [9]	ERT [11]	LBF [8]	CGPRT [15]	CFSS [14]	GSDM [10]	CSDM	CSDMa
300W	7.58	8.38	7.52	6.40	6.32	**5.71**	5.76	6.96	6.83	6.40
BU4DFE-S	9.45	8.61	9.57	-	-	15.81	-	9.01	8.28	**7.62**

Table 1: Our method compared with state-of-the-art methods in terms of mean landmark displacement as percentage of interocular distance[2] without (CSDM) and with multiple test initialisations (CSDMa).

pitch and yaw, this dataset better represents the strength of the proposed algorithm to better adjust to the main modes of variation of the data. This is analysed in Figure 3. There, the NMEE is shown relative to the yaw rotation, for two ranges of pitch. Without using multiple test initialisations, the proposed method has an accuracy similar to that of the other state-of-the-art approaches for near-frontal faces, but is much more robust to pose variations. It works specially well for both large pitch and yaw rotations. This contrasts with CGPRT, which performed specially well for the 300W dataset, but had problems with BU4DFE-S. The only method still far from, but approaching the accuracy obtained by the proposed approach is RCPR. It can be seen in Figure 3 that while RCPR has the lowest NMEE for frontal faces, it one of the best approaches when dealing with large pose variations. When using multiple test initialisations, a much lower average error is obtained, achieving the same accuracy for near-frontal faces as ERT. This accuracy improvement is maintained regardless of the facial pose, except for large pose rotations in both pitch and yaw, where the yaw angle is close to 90°. For these extreme cases, the error is only slightly lower than CSDM without using multiple shape initialisations.

A breakdown of the NMEE by facial regions, as shown in Table 2, gives a better insight on the method performance. For far from frontal head poses, the proposed approach surpasses the state of the art accuracies on all facial regions, both with and without multiple test initialisations. In the case of near-frontal head poses, RCPR has a higher precision for the eyes and eyebrows. CSDM is better at localising landmarks at the nose, mouth and contour regions when using multiple shape initialisations. An interesting result is the error reduction when localising the contour landmarks with multiple shape initialisations. While the other facial regions reduce the RMSE by about 5%, in the case of the contour it is reduced by over 10%, both in close to and far from frontal head poses. This is likely caused by the lack of edges and strong gradients on this region. By averaging multiple predictions, the noise is reduced, obtaining a higher accuracy.

GSDM is another method that exploits the features main modes of variation to better approximate the descent direction at different regions of the feature space. Compared to it, the proposed method obtains better results for both 300W and BU4DFE-S while also producing a more compact model. The memory required by GSDM increases quadratically with the number of considered bases, while the proposed approach does so linearly. Furthermore, each position of the subspace has a unique regressor assigned, while GSDM shares the same regression weights for a given partition of the subspace. One downside to the

	Close to frontal							Far from frontal						
	ESR	RCPR	SDM	CGPRT	GSDM	CSDM	CSDMa	ESR	RCPR	SDM	CGPRT	GSDM	CSDM	CSDMa
eyes	3.92	**3.38**	4.02	10.53	3.92	4.04	3.82	6.94	6.11	6.76	14.29	6.25	5.55	**5.20**
eyebrows	5.84	**5.17**	5.60	13.15	5.56	5.84	5.54	9.01	8.02	8.50	17.73	8.12	7.20	**6.77**
nose	6.03	5.59	5.60	10.30	5.51	5.58	**5.27**	8.26	7.69	8.58	13.21	8.00	7.41	**6.99**
mouth	5.46	4.28	4.47	10.91	4.27	4.40	**4.27**	8.20	6.70	8.18	14.52	6.72	6.17	**5.84**
contour	12.59	12.11	13.26	17.49	13.52	13.27	**12.04**	17.30	17.19	18.54	22.43	18.53	17.20	**15.27**

Table 2: Normalised Mean Euclidean Error (NMEE) for different landmark subsets corresponding to facial regions on BU4DFE-S. We group faces according to their pose. Close-to-frontal faces have an yaw angle between ±30° and pitch angle between ±15°. Correspondingly far-from frontal faces have both yaw and pitch angles above ±30° and ±15° respectively.

proposed approach is that the computational cost increases linearly with the number of bases, while for GSDM the cost remains constant.

Similarly to SDM and GSDM, the proposed method provides a closed-form solution. Compared to other state-of-the-art methods such as CFSS, CGPRT and LBF, which use stochastic processes when learning each regressor, the proposed approach ensures a consistent result on different training runs given the same data.

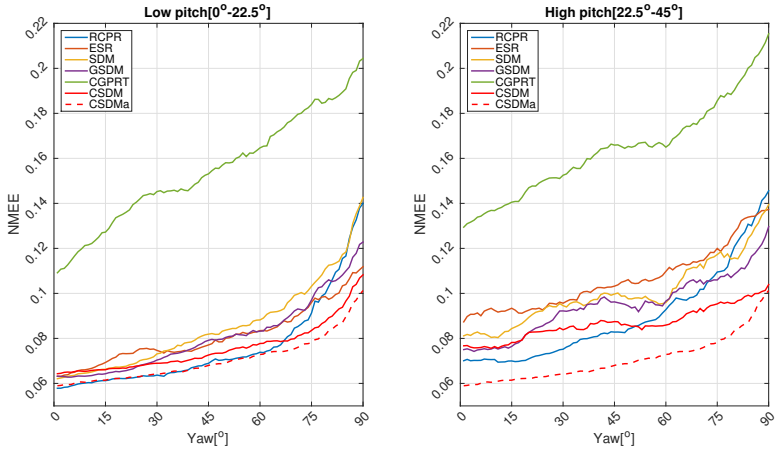


Fig. 3: Normalised Mean Euclidean Error (NMEE) as a function of yaw on two different pitch ranges on BU4DFE-S.

Multiple qualitative examples of faces from the BU4DFE-S dataset, with the landmark predictions for different methods, are shown in Figure 4. From these examples it can be seen that SDM, CGPRT and RCPR struggle to correctly locate inner face landmarks for heavily rotated faces. Compared to all other considered methods, our proposal has a high accuracy on inner face landmarks

even with highly rotated faces, followed by GSDM and ESR. The main weakness is the localisation of face contour landmarks, which is noisy due to the lack of edges and little texture information on that area, resulting in a lack of smoothness in the contour line. Even with this noise, as shown in Table 2, the proposed approach has a much better precision for this set of landmarks. An extension to consider in the future would be regressing a parametrised shape, which should increase the accuracy for the face contours.

4 Conclusion

In this work we extended cascaded regression approaches by introducing the second order derivative over the main modes of variation of the features, presenting a closed-form solution to the face alignment problem. We showed that by doing so, the robustness to large head pose variations is greatly increased, surpassing current state of the art methods. At the same time, the accuracy for near-frontal faces is comparable to state of the art results. Furthermore, the learnt models are smaller than those from other similar approaches.

In order to prove the effectiveness of our method on heavily rotated faces we have built a new synthetic dataset based on a well known public 3D face dataset. It contains large variations in both head pose and facial expressions, as well as a large number of training instances, making it one of the largest, most challenging datasets for facial landmark localisation to date.

Several future improvements can be envisioned, like parameterizing the face to increase shape consistency especially for landmarks situated in regions with little texture and extending the method to 3D, which would make it useful for a larger number of applications.

Acknowledgement. The work of Marc Oliu is supported by the FI-DGR 2016 fellowship, granted by the Universities and Research Secretary of the Knowledge and Economy Department of the Generalitat de Catalunya. This work has been partially supported by the Spanish project TIN2013-43478-P, the European Comission Horizon 2020 granted project SEE.4C under call H2020-ICT-2015 and the U.S. National Institutes of Health under the grant MH096951.

References

1. Corneanu, C.A., Oliu, M., Cohn, J.F., Escalera, S.: Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. Transactions on Pattern Analysis and Machine Intelligence Special Issue (2016)
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis & Machine Intelligence (2001) 681–685
3. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding **61** (1995) 38–59

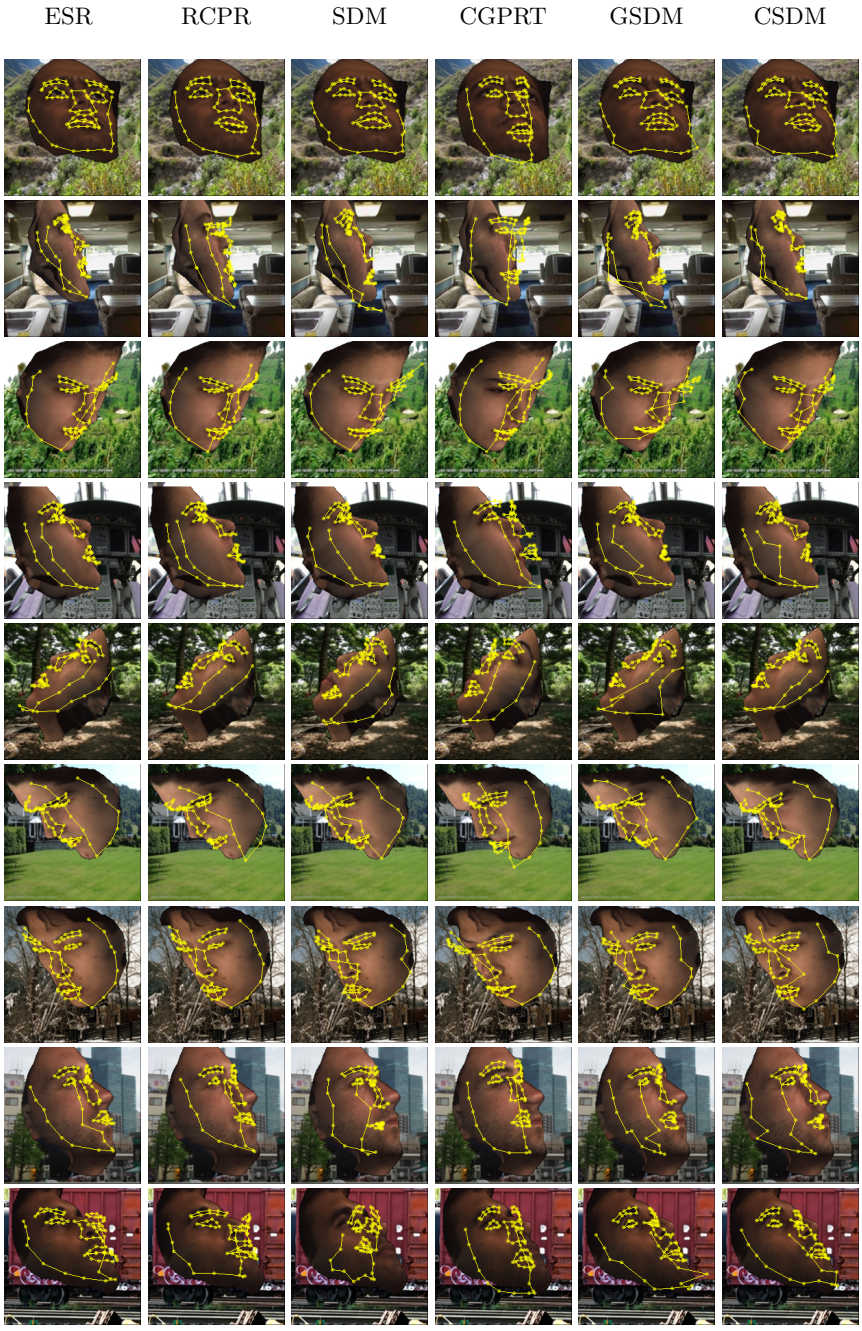


Fig. 4: Facial landmark localisation examples for BU4DFE-S.

4. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3476–3483
5. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: Computer Vision–ECCV 2014. Springer (2014) 1–16
6. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision **107** (2014) 177–190
7. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1513–1520
8. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1685–1692
9. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 532–539
10. Xiong, X., De la Torre, F.: Global supervised descent method. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2664–2673
11. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1867–1874
12. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3d face alignment from 2d videos in real-time. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. Volume 1., IEEE (2015) 1–8
13. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 1078–1085
14. Zhu, S., Li, C., Change Loy, C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 4998–5006
15. Lee, D., Park, H., Yoo, C.D.: Face alignment using cascade gaussian process regression trees. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 4204–4212
16. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, Ieee (2012) 510–517
17. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2013) 397–403
18. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2879–2886
19. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35** (2013) 2930–2940
20. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Computer Vision–ECCV 2012. Springer (2012) 679–692

21. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On, IEEE (2008) 1–6
22. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Advances in neural information processing systems. (2014) 487–495
23. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28** (2010) 807–813