RAPS: Robust and Efficient Automatic Construction of Person-Specific Deformable Models

Christos Sagonas*, Yannis Panagakis*, Stefanos Zafeiriou*, and Maja Pantic*†

*Department of Computing,
Imperial College London,
180 Queens Gate,
London SW7 2AZ, U.K.

†EEMCS,
University of Twente,
Drienerlolaan 5,
7522 NB Enschede, The Netherlands

{c.sagonas, i.panagakis, s.zafeiriou, m.pantic}@imperial.ac.uk

Abstract

The construction of Facial Deformable Models (FDMs) is a very challenging computer vision problem, since the face is a highly deformable object and its appearance drastically changes under different poses, expressions, and illuminations. Although several methods for generic FDMs construction, have been proposed for facial landmark localization in still images, they are insufficient for tasks such as facial behaviour analysis and facial motion capture where perfect landmark localization is required. In this case, person-specific FDMs (PSMs) are mainly employed, requiring manual facial landmark annotation for each person and person-specific training.

In this paper, a novel method for the automatic construction of PSMs is proposed. To this end, an orthonormal subspace which is suitable for facial image reconstruction is learnt. Next, to correct the fittings of a generic model, image congealing (i.e., batch image aliment) is performed by employing only the learnt orthonormal subspace. Finally, the corrected fittings are used to construct the PSM. The image congealing problem is solved by formulating a suitable sparsity regularized rank minimization problem. The proposed method outperforms the state-of-the art methods that is compared to, in terms of both landmark localization accuracy and computational time.

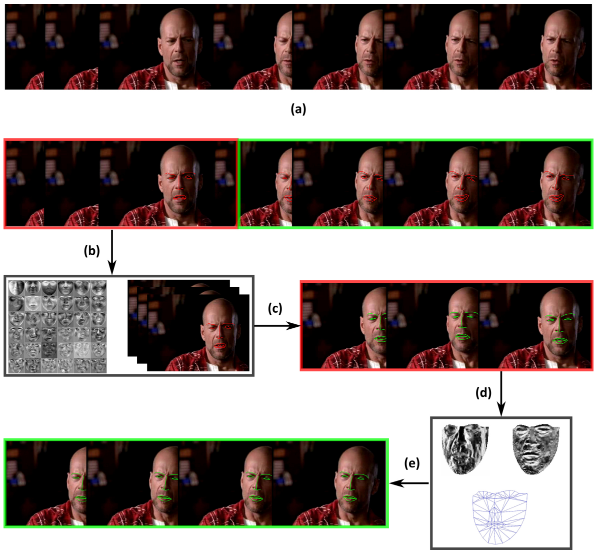


Figure 1. **Automatic construction of a Person-Specific Model**. (a) Frames of a video sequence depict the same person. (b) Results from a facial landmark detector. The proposed method takes a small number of initial frames with the corresponding erroneous initializations, and finds the corrections (c) by using the PCA bases of a different face database. Furthermore, a Person Specific Model is built using the corrected frames (d) and the rest frames of the video are fitted (e).

1. Introduction

The construction of generic models, which are able to capture the variability of deformable objects is one of the most popular and well-studied computer vision problems. Arguably, the most studied deformable object is the human face [11].

The methods employed for the construction of Facial Deformable Models (FDMs) are roughly classified into two categories, based on whether they use the entire face region or local image patches. In particular, the *holistic methods*, such as the Active Appearance Models (AAMs) [11, 19, 20, 27]) employ a holistic texture-based facial representation. In contrast, the *parts-based methods* represent the face via a set of local image patches cropped around the landmark points. The most notable examples of the latter category are the Active Shape Models (ASMs) and Constrained Local Models (CLMs) [2, 12, 13, 25]. It is worth mentioning that, methods such as those in [3, 7, 29], are not strictly fall within the aforementioned two categories. Furthermore, both the construction of the generic FDM, as well as, the optimization strategy employed to fit the FDMs in unseen (test) images can be also divided into two categories, namely the *generative* [11, 19, 20, 27] and the *discriminative* [2, 3, 18]. This categorization is done based on whether the methods use discriminative information (i.e., a set of facial landmark classifiers [2, 25] or a discriminative cost function [18]) or a generative analysis-by-synthesis approach [11, 19, 20, 27].

Many of the aforementioned generic FDMs have successfully been applied in facial landmark localization in still images by employing *in-the-wild* experimental scenarios [2, 27, 29]. This is not the case for in-the-wild facial feature tracking, where appropriate annotated datasets are not available yet. Without exception, the above mentioned methods rely on a static generic model that is trained completely on off-line training data. Nevertheless, when it comes to applications that require perfect facial landmark localization and tracking accuracy, such as the analysis of human facial behavior (e.g., Facial expressions and Facial Action Unit (FAU) recognition [10]), as well as, facial motion capture, generic models are insufficient. To this end, person-specific models (PSMs) are mainly applied [10, 28]. However, the construction of PSMs requires manually annotation of images depicting the person, which is a laborious and time consuming process. Consequently, the automatic construction of PSMs is of paramount importance.

Various method for the automatic construction of PSMs have been proposed [26]. These methods apply incremental subspace learning, such as the incremental Principal Component Analysis (iPCA) [17] onto a set of fittings produced by a generic AAM in order to update the model. The drawback of blindly applying incremental subspace learning without incorporating a correction strategy is that, erroneous fittings may arbitrarily bias the learnt subspace, resulting in model drifting.

In this paper, a novel method for the automatic construction of PSMs is proposed, aiming at alleviating the drawbacks caused by the erroneous fittings. The method is referred to as *Robust and Efficient Automatic Construction of Person-Specific Deformable Models* (RAPS). To this end, we first learn an orthonormal subspace which is suitable for facial image reconstruction, by using manually annotated data which have been collected in-the-wild. Next, we perform image congealing (i.e., batch image aliment) using the fittings of a generic model in order to correct them. This procedure is performed by employing only the learnt orthonormal subspace. Finally, the corrected fittings are used for the construction of the PSM. In RAPS, the image congealing is performed by solving a suitable sparsity regularized rank minimization problem. Compared to the related image congealing methods in [8, 9, 21, 30], the RAPS not only avoids the unnatural deformations, since the faces always lie in the face subspace, but also it has lower computational complexity.

To assess the performance of the RAPS, experiments on image congealing and person specific facial modelling have been conducted. The experimental results indicate that, the RAPS outperforms the state-of-the-art methods [8, 21, 30] in terms of landmark localization accuracy and computational time.

2. Correcting Erroneous Fittings using a Point Distributional Model

In this section, some preliminaries regarding the alignment of AAMs are briefly summarized. Next, recent methods for erroneous fittings correction using a shape model and rank minimization are briefly reviewed.

The AAMs employ statistical models to describe the variations of shape and texture. In particular, a set of annotated points is used to learn a statistical model of shape. To retain only the variability that is attributed to non-rigid deformations, the shapes are put in correspondence, usually by removing the global similarity transforms [20] via a Procrustes analysis. Similarly, a statistical model of the texture is learnt using textures that are in correspondence with respect to the shape points (i.e., the so-called shape-free textures). This requires a predefined reference frame, usually defined by the mean shape, and a global motion model (i.e., the warp, namely piece-wise affine or Thin-Plate Spline model). The two main assumptions behind AAMs are: 1) for every unseen (test) texture there exists a set of weights, allowing the warped test image into the mean shape to be written as a linear combination of the shape-free texture model plus the mean frame and 2) the test shape can be written as a linear combination of the training shapes. In mathematical terms, let us consider an L landmark shape model and a reference frame of F pixels, then let $\mathcal{S} = \{\bar{\mathbf{s}}, \mathbf{B} \in \Re^{2L \times p}\}$ and $\mathcal{T} = \{\bar{\mathbf{x}}, \mathbf{U} \in \Re^{F \times m}\}$ be the linear models for the shape and texture, respectively. The bases of the shape $\mathbf{B}$ and of texture $\mathbf{U}$ are computed using Principal Component Analysis (PCA). Also, for simplicity, the global similarity transformations are incorporated as additional 4 bases in $\mathbf{B}$. The updated $\mathbf{B}$ is used in the rest of the paper. Then, according to the above assumptions for a test shape $\mathbf{s} \in \Re^{2L}$ of a test image $\mathbf{x}$ we have the approximation:

$$\begin{aligned} \mathbf{s} &\approx \bar{\mathbf{s}} + \mathbf{B}\mathbf{p} \\ \mathbf{x}(W(\mathbf{p})) &\approx \bar{\mathbf{x}} + \mathbf{U}\mathbf{c}, \end{aligned} \tag{1}$$

where $\mathbf{x}(W(\mathbf{p}))$ is the vectorized warped test image in the reference frame (from now onwards for simplicity reasons instead of $\mathbf{x}(W(\mathbf{p}))$ we will use $\mathbf{x}(\mathbf{p})$). Under the above assumptions, the parameters $\mathbf{p}, \mathbf{c}$ are computed by minimizing the error of the reconstruction of the shape-free texture, using the statistical texture model:

$$\mathbf{p}_o, \mathbf{c}_o = \min_{\mathbf{p},\mathbf{c}} \|\mathbf{x}(\mathbf{p}) - (\bar{\mathbf{x}} + \mathbf{U}\mathbf{c})\|^2_{\mathbf{P}}. \tag{2}$$

Where $\mathbf{P}$ are appropriate projection operators and $\|\mathbf{x}\|^2_{\mathbf{P}} = \mathbf{x}^T\mathbf{P}\mathbf{x}$. The solution of the above optimization problem is referred to as model fitting. Many optimization methods have been proposed to fit the model in test images [11, 20]. The most popular are the regression-based fitting [11] and the Project-out Inverse Compositional (PIC) method [20]. Robust methods to fit AAM include the approach [1].

Problem formulation. In this paper, the following problem is investigated. We assume, that we have a generic face tracker, such as a generic AAM or a CLM which has been applied to a number of frames of a video sequence of a persons' face. We want to use the set of N erroneous fittings $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]$ to build an accurate PSM that can be used to track the rest of the video.

Notations. In the following, the rank$(\mathbf{X})$ is the rank of matrix $\mathbf{X}$ (i.e., the maximum number of linearly independent row or column vectors of $\mathbf{X}$). The matrix ℓ_0 quasi-norm is denoted by $\|\mathbf{X}\|_0$ and returns the number of nonzero entries in $\mathbf{X}$. The matrix ℓ_1 norm is defined as $\|\mathbf{X}\|_1 = \sum_i \sum_j |x_{ij}|$, where $| \cdot |$ denotes the absolute value operator. The Frobenius norm is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j x^2_{ij}}$. The nuclear norm of $\mathbf{X}$ (i.e., the sum of singular values of a matrix) is denoted by $\|\mathbf{X}\|_*$. The ℓ_2 norm of a vector $\mathbf{x}$ is denoted as $\|\mathbf{x}\|_2$.

2.1. Robust Alignment by Sparse and Low-ran Decomposition, RASL

The first method proposed for the problem under study was published in [21]. The rationality behind the method in [21] is that, given a collection $\mathbf{X}$ of images that lie in a low-rank space, (e.g., facial images of one person) and a set of initialization parameters $\{\mathbf{p}^c_i\}^N_{i=1}$, it is possible to simultaneously estimate the low rank subspace $\mathbf{A}$ and the alignment parameters. Furthermore, by incorporating an error matrix $\mathbf{E}$, the procedure can be highly robust to outliers. Formally, the problem is to find the increments $\{\delta\mathbf{p}_i\}^n_{i=1}$ which are being applied in matrix $\mathbf{X}$ as $\mathbf{X}(\{\delta\mathbf{p}_i\}^N_{i=1}) = [\mathbf{x}(W(\mathbf{p}^c_1 + \delta\mathbf{p}_1)), \dots, \mathbf{x}(W(\mathbf{p}^c_n + \delta\mathbf{p}_N))]$ such that the matrix $\mathbf{A}$ to be of low-rank and the error matrix to be sparse. That is, to solve:

$$\begin{aligned}&\min\nolimits_{\mathbf{A},\mathbf{E},\{\delta p_i\}^N_{i=1}} \text{rank}(\mathbf{A}) + \lambda\|\mathbf{E}\|_0,\\ &\text{s.t. } \mathbf{X}(\{\mathbf{p}^c_i + \delta\mathbf{p}_i\}^N_{i=1}) = \mathbf{A} + \mathbf{E},\end{aligned} \tag{3}$$

which is an NP-hard problem due to the presence of rank operator and the ℓ_0 quasi-norm. A convex relaxation of (3) can be formulated by replacing the rank operator and the ℓ_0 quasi-norm in with their convex surrogates, namely the nuclear norm and the ℓ_1 norm, respectively. Furthermore, the linearization of image $\mathbf{x}_i(W(\mathbf{p}^c_i + \delta\mathbf{p}_i))$ around $\mathbf{p}^c_i$ as:

$$\mathbf{x}_i(\mathbf{p}^c_i + \delta\mathbf{p}_i) \approx \mathbf{x}_i(\mathbf{p}_i) + \mathbf{J}_{\mathbf{x}}|_{\mathbf{p}=\mathbf{p}^c_i}\delta\mathbf{p}_i, \tag{4}$$

where $\mathbf{J}_{\mathbf{x}_i} \in \Re^{F\times p}$ is the image Jacobian expanded by the chain rule as $\mathbf{J}_{\mathbf{x}_i} = \nabla_W \mathbf{x}_i \frac{\partial W}{\partial \mathbf{p}}$, will be used. The symbol $|$ indicates where the Jacobian is computed. Details the computation of $\frac{\partial W}{\partial \mathbf{p}}$ are provided in [20]. Consequently, a convex relaxation of (3) is:

$$\begin{aligned}&\min\nolimits_{\mathbf{A},\mathbf{E},\{\delta\mathbf{p}_i\}^N_{i=1}} \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1,\\ &\text{s.t. } \mathbf{X}(\{\mathbf{p}^c_i\}^N_{i=1}) + \textstyle\sum^n_{i=1} \mathbf{J}_{\mathbf{x}_i}\delta\mathbf{p}_i\hat{\mathbf{e}}^T_i = \mathbf{A} + \mathbf{E}\end{aligned} \tag{5}$$

where $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_n$ is the standard bases of $\Re^n$. Even though in [21] only simple motion models, such as global affine, were consider, the application of statistical shape model is straightforward. The computational complexity of (5) is that of the Singular Value Decomposition (SVD) i.e., $\mathcal{O}(l(\min(F, N)^3 + N^2F))$, where l is the total number of iterations required for convergence.

2.2. RASL using Anchor Points, A-RASL

In [8, 9], it was argued that if complex motion models, such as a piece-wise affine motion model driven by a point distributional model, are incorporated into the RASL, the subject's facial appearance in the image ensemble will be deformed arbitrarily. Consequently, a false alignment will be obtained. To alleviate this problem, it was proposed to incorporate the so-called anchor shapes [8, 9]. The anchor shapes are just original shapes of a generic tracker $\{\mathbf{s}^a\}^N_{i=1}$, which are used to penalize the arbitrary warping of face appearance. To this end, the following optimization problem is solved:

$$\begin{aligned}&\min\nolimits_{\mathbf{A},\mathbf{E},\{\delta\mathbf{p}_i\}^N_{i=1}}\\ &\|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 + \lambda_1 \textstyle\sum^N_{i=1}\sum^L_{j=1} \|\mathbf{z}_{ij}\|_2,\\ &\text{s.t. } \mathbf{X}(\{\mathbf{p}^i_c\}^N_{i=1}) + \textstyle\sum^N_{i=1} \mathbf{J}_i\delta\mathbf{p}_i\hat{\mathbf{e}}^T_i = \mathbf{A} + \mathbf{E}\\ &\mathbf{z}_i = \mathbf{s}(\mathbf{p}^i_c) + \mathbf{\Phi}(\mathbf{p}^i_c)\delta\mathbf{p}_i - \mathbf{s}^a_i\end{aligned} \tag{6}$$

where $\mathbf{z}_{ij} = [x_{ij}\ y_{ij}]$ is the vector of coordinates for the j point and $\mathbf{\Phi}(\mathbf{p}^i_c)$ are the shape Jacobians.

The drawbacks of this approach are: 1) since it penalizes huge changes the solution is always close to the original erroneous fittings and 2) there is an additional regularization term, for which the weight parameter λ_1 is difficult to be tuned. Furthermore, the computational complexity of the method is the same as that of the RASL.

2.3. Using a generic model for regularization, GMR

In [30], it is proposed to use also a generic statistical model to regularize the simultaneous alignment and subspace estimation procedure. That is, instead of using the anchor points to regularize the shapes, a generic statistical model of a face is used to regularize the texture and in that way the textures in matrix $\mathbf{X}$ cannot be arbitrarily warped. The general statistical model employed, is a matrix of bases $\mathbf{U}$ obtained via PCA, as in the case of AAMs.

More formally, the rationality behind [30] is that a set of alignment parameters $\{\delta \mathbf{p}_i\}_{i=1}^N$, such that the aligned images are as close as possible to the generic model $\mathbf{U}$ and the aligned images lie in a low-rank subspace, i.e., $\mathbf{A}$ can be found by solving:

$$\begin{array}{rl} \min_{\{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{C}} & \|\mathbf{A}\|_* + \lambda \|\mathbf{A} - \mathbf{M} - \mathbf{U}\mathbf{C}\|_F^2 \\ \text{s.t.} & \mathbf{A} = \mathbf{X}(\{\mathbf{p}_i^c\}_{i=1}^N) + \sum_{i=1}^N \mathbf{J}_i \delta \mathbf{p}_i \hat{\mathbf{e}}_i^T, \end{array} \tag{7}$$

where $\mathbf{M}$ is the matrix composed of replicates of $\bar{\mathbf{x}}$. The method is not expected to be as robust as the RASL is in case of occlusions-outliers, since the Frobenius norm is used for error handling. The computational complexity of (7) is again the same as that of the RASL.

2.4. The proposed approach

In this paper, we make use of a generic model, not as regularization term in as in [30], but in order to decrease the computational complexity. Since we restrict ourselves only in the generic clean face subspace $\mathbf{U}$ the optimization problem is written as:

$$\begin{array}{rl} \min_{\{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{C}, \mathbf{E}} & \text{rank}(\mathbf{U}\mathbf{C}) + \lambda \|\mathbf{E}\|_0 \\ \text{s.t.} & \mathbf{X}(\{\mathbf{p}_i^c + \delta \mathbf{p}_i\}_{i=1}^N) = \mathbf{U}\mathbf{C} + \mathbf{M} + \mathbf{E}, \end{array} \tag{8}$$

where $\lambda > 0$. By replacing the rank operator with the nuclear norm and the ℓ_0-norm with ℓ_1-norm, (8) is written as:

$$\begin{array}{rl} \min_{\{\mathbf{p}_i\}_{i=1}^N, \mathbf{C}, \mathbf{E}} & \|\mathbf{U}\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_1, \\ \text{s.t.} & \mathbf{X}(\{\mathbf{p}_i + \delta \mathbf{p}_i\}_{i=1}^N) = \mathbf{U}\mathbf{C} + \mathbf{M} + \mathbf{E}. \end{array} \tag{9}$$

Since $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, and the unitary invariance of the nuclear norm, (9) is equivalent to:

$$\begin{array}{rl} \min_{\{\delta \mathbf{p}_i^c\}_{i=1}^N, \mathbf{C}, \mathbf{E}} & \|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_1 \\ \text{s.t.} & h(\mathbf{C}, \{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{E}) = 0, \end{array} \tag{10}$$

where,

$$h(\mathbf{C}, \{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{E}) = \mathbf{X}(\{\mathbf{p}_i + \delta \mathbf{p}_i\}_{i=1}^N) - \mathbf{U}\mathbf{C} - \mathbf{M} - \mathbf{E}. \tag{11}$$

Clearly, (10) involves the minimization of the nuclear norm of a matrix $\mathbf{C} \in \Re^{k \times N}$, where $k \ll F$.

Problem (10) can be solved iteratively by the *augmented Lagrange multiplier method* (ALM) [4]. That is, (10) is solved by minimizing the augmented Lagrangian function:

$$\begin{aligned} &\mathcal{L}(\{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{C}, \mathbf{E}, \mathbf{\Lambda}) \\ &= \|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_1 + \text{tr}\left(\mathbf{\Lambda}^T (h(\mathbf{C}, \{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{E}))\right) \\ &+ \frac{\mu}{2} \|h(\mathbf{C}, \{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{E})\|_F^2, \end{aligned} \tag{12}$$

where $\mathbf{\Lambda}$ are the Lagrange multipliers for the equality constraints in (10) and μ is a non-negative penalty parameter. By employing the ALM, (12) is minimized with respect to each variable in an alternating fashion and finally the Lagrange multipliers are updated at each iteration as outlined in Algorithm 1.

Since $\mathbf{U}$ is an orthonormal matrix, the trace part of (12) is written as:

$$\begin{array}{c} \text{tr}\left(\mathbf{\Lambda}^T (h(\mathbf{C}, \{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{E}))\right)_{\mathbf{I} - \mathbf{U}\mathbf{U}^T} \\ + \text{tr}\left(\mathbf{\Lambda}^T (h(\mathbf{C}, \{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{E}))\right)_{\mathbf{U}\mathbf{U}^T}, \end{array} \tag{13}$$

which is equal to:

$$\begin{array}{l} \text{tr}\left[\mathbf{\Lambda}^T (\mathbf{I} - \mathbf{U}\mathbf{U}^T)(\mathbf{X}\{\mathbf{p} + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E})\right] + \\ \text{tr}\left[(\mathbf{U}^T\mathbf{\Lambda})^T (\mathbf{U}^T (\mathbf{X}\{\mathbf{p} + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E}) - \mathbf{C})\right]. \end{array} \tag{14}$$

Similarly, $\|h(\mathbf{C}, \{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{E})\|_F^2$ can be rewritten as:

$$\begin{array}{l} \|\mathbf{X}\{\mathbf{p}_i + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E}\|_{\mathbf{I} - \mathbf{U}\mathbf{U}^T}^2 + \\ \|\mathbf{U}^T (\mathbf{X}\{\mathbf{p}_i + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E}) - \mathbf{C}\|_{\mathbf{U}\mathbf{U}^T}^2. \end{array} \tag{15}$$

Consequently, $\mathbf{C}$ is found by minimizing:

$$\begin{array}{l} \|\mathbf{C}\|_* + \\ \text{tr}\left[(\mathbf{U}^T\mathbf{\Lambda})^T (\mathbf{U}^T (\mathbf{X}\{\mathbf{p} + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E}) - \mathbf{C})\right] + \\ \frac{\mu}{2} \|\mathbf{U}^T (\mathbf{X}\{\mathbf{p}_i + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E}) - \mathbf{C}\|_{\mathbf{U}\mathbf{U}^T}^2. \end{array} \tag{16}$$

The error matrix $\mathbf{E}$ is the solution of the following optimization problem:

$$\min_{\mathbf{E}} \mathcal{L}(\{\delta \mathbf{p}_i\}_{i=1}^N, \mathbf{C}, \mathbf{E}, \mathbf{\Lambda}). \tag{17}$$

Problems (16) and (17) are solved by employing proximal operators. In particular, (16) is solved by the singular value thresholding operator (SVT) defined for any matrix $\mathbf{Q}$ as [5]: $\mathcal{D}_\tau[\mathbf{Q}] = \mathbf{U}\mathcal{S}_\tau \mathbf{V}^T$ with $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ being the singular value decomposition and $\mathcal{S}_\tau[q] = \text{sgn}(q)\text{max}(|q| - \tau, 0)$ is the shrinkage operator [6]. (17) is solved by the shrinkage operator by applying it element-wise.

Finally, the parameters' increments $\delta \mathbf{p}_i$ are obtained by minimizing:

$$\begin{array}{l} \text{tr}\left[\mathbf{\Lambda}^T (\mathbf{I} - \mathbf{U}\mathbf{U}^T)(\mathbf{X}\{\mathbf{p}_i + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E})\right] + \\ \frac{\mu}{2} \|\mathbf{X}\{\mathbf{p}_i + \delta \mathbf{p}_i\}_{i=1}^N - \mathbf{M} - \mathbf{E}\|_{\mathbf{I} - \mathbf{U}\mathbf{U}^T}^2. \end{array} \tag{18}$$

That is, the solution of (18) (i.e., the increments $\delta\mathbf{p}_i$ for each image $\mathbf{x}_i$) is given by:

$$\delta\mathbf{p}_i = -[\tilde{\mathbf{J}}_{x_i}^T \tilde{\mathbf{J}}_{x_i}]^{-1} \mathbf{J}_{x_i}^T (\mathbf{X}\{\mathbf{p}_i\} - \bar{\mathbf{x}} - \mathbf{e}_i - \mathbf{l}_i/\mu), \quad (19)$$

where $\mathbf{e}_i$ is the i column of matrix $\mathbf{E}$, $\mathbf{l}_i$ is the i column of the Lagrange multipliers matrix $\mathbf{\Lambda}$, $\mathbf{J}_{x_i}^T$ is the Jacobian of image $\mathbf{x}_i$ and $\tilde{\mathbf{J}}_{x_i}^T$ is the Jacobian projected at $\mathbf{I} - \mathbf{U}\mathbf{U}^T$. In order to calculate efficiently the term $\tilde{\mathbf{J}}_{x_i}^T \tilde{\mathbf{J}}_{x_i}$ we use the formulation below:

$$\begin{aligned} \tilde{\mathbf{J}}_{x_i}^T \tilde{\mathbf{J}}_{x_i} &= \tilde{\mathbf{J}}_{x_i}^T (\mathbf{I} - \mathbf{U}\mathbf{U})^T \mathbf{J}_{x_i}^T = \\ &= \tilde{\mathbf{J}}_{x_i}^T \mathbf{J}_{x_i} - (\mathbf{U}^T \tilde{\mathbf{J}}_{x_i})^T (\mathbf{U}^T \mathbf{J}_{x_i}). \end{aligned} \quad (20)$$

The computational complexity of Algorithm 1 is that of the SVD involved in the computation of SVT i.e. $\mathcal{O}(l(\min(F,N)^3 + N^2 k))$, where k is the number of bases in $\mathbf{U}$ and l is the total number of iterations.

3. Experimental Evaluation

The aims of the experiments are: 1) to show that the proposed method can be used for effective and efficient image congealing using images captured both in constrained and unconstrained conditions and most importantly, 2) to show that the shapes produced by the RAPS can be used to build a robust PSM.

The performance of the RAPS is compared against of that obtained by the three state-of-the-art methods, namely the RASL [21], the A-RASL [8], and the GMR [30]. It should be noted that the RASL employs only rigid affine transformations and thus it cannot be compared with the RAPS, the A-RASL, and the GMR. To this end, an enhanced versions of RASL, referred to as eRASL, was implemented. The eRASL employs a shape model with a piecewise affine motion model.

The initial fittings were provided by an in-house version of the face and landmark detector proposed in [31]. The shape model $\mathcal{S} = \{\bar{\mathbf{s}}, \mathbf{B}\}$ for all tested methods, as well as the appearance model $\mathcal{T} = \{\bar{\mathbf{x}}, \mathbf{U}\}$ used in the GMR and the RAPS, were learnt from the publicly available in-the-wild database AFW [31]. The 68-landmark annotations of the AFW which used for the construction of $\mathcal{S}$, and $\mathcal{T}$, were retrieved from [23]. In the implementation of the RAPS, the weight parameter λ was set as $\lambda = 1/N$, where N is the number of images from the same subject. The number of bases (k) in $\mathbf{U}$ was set 180.

Algorithm 1 Solving (12) by the ALM method.

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{F \times N}$, PCA bases $\mathbf{U} \in \mathbb{R}^{F \times k}$, and the parameter λ.
Output: Matrix $\mathbf{X}\{\mathbf{p}\} \in \mathbb{R}^{F \times N}$.

```
while not converged do
    Compute Jacobian matrices.
    Warp and normalize the images.
    Initialize: $\mathbf{C}_{[0]} = \mathbf{0}, \mathbf{E}_{[0]} = \mathbf{0}, \mathbf{\Lambda}_{[0]} = \mathbf{0}, \mu_{[0]} = 10^{-6}, \rho = 1.2, \epsilon_1 = 10^{-8}$, and $\epsilon_2 = 10^{-5}$.
    while not converged do
        Fix the other variables and update $\mathbf{C}_{[t+1]}$ by:
        $\mathbf{C}_{[t+1]} \leftarrow \min_{\mathbf{C}_{[t]}} \mathcal{L}(\{\delta\mathbf{p}_{i,[t]}\}_{i=1}^N, \mathbf{C}_{[t]}, \mathbf{E}_{[t]}, \mathbf{\Lambda}_{[t]})$
            $= \mathcal{D}_{\frac{1}{\mu_{[t]}}} \left[\mathbf{U}^T \left(\mathbf{X} - \mathbf{E}_{[t]} - \mathbf{M} + \mathbf{\Lambda}_{[t]}/\mu_{[t]}\right)\right]$
        Fix the other variables and update $\mathbf{E}_{[t+1]}$ by:
        $\mathbf{E}_{[t+1]} \leftarrow \min_{\mathbf{E}_{[t]}} \mathcal{L}(\{\delta\mathbf{p}_{i,[t]}\}_{i=1}^N, \mathbf{C}_{[t+1]}, \mathbf{E}_{[t]}, \mathbf{\Lambda}_{[t]})$
            $= \mathcal{S}_{\frac{\lambda}{\mu_{[t]}}} \left[\mathbf{X} - \mathbf{U}\mathbf{C}_{[t+1]} - \mathbf{M} + \mathbf{\Lambda}_{[t]}/\mu_{[t]}\right]$
        Fix the other variables and update $\delta\mathbf{p}_i$ for the image $i$ by:
            $\delta\mathbf{p}_{i,[t+1]} \leftarrow \left[\tilde{\mathbf{J}}_{x_i}^T \mathbf{J}_{x_i}{}^T\right]^{-1} \mathbf{J}_{x_i}{}^T$
                $(\mathbf{X} - \bar{\mathbf{x}} - \mathbf{i}_{e,[t+1]} - \mathbf{l}_{i,[t]}/\mu_{[t]})$
        Update the Lagrange multiplier by:
        $\mathbf{\Lambda}_{[t+1]} \leftarrow \mathbf{\Lambda}_{[t]} + \mu_{[t]}(h(\mathbf{C}_{[t+1]}, \{\delta\mathbf{p}_{i,[t+1]}\}_{i=1}^N, \mathbf{E}_{[t+1]}))$.
        Update $\mu_{[t+1]}$ by $\mu_{[t+1]} \leftarrow \min(\rho \cdot \mu_{[t]}, 10^{10})$.
        Check convergence conditions:
        $\|\mathbf{X} - \mathbf{U}\mathbf{C}_{[t+1]} - \mathbf{M} - \mathbf{E}_{[t+1]}\|_F / \|\mathbf{X}\|_F \leq \epsilon_1$
        and
        $\max\left(\|\mathbf{E}_{[t]} - \mathbf{E}_{[t-1]}\|_F / \|\mathbf{X}\|_F,\right.$
        $\left.\|\mathbf{C}_{[t]} - \mathbf{C}_{[t-1]}\|_F / \|\mathbf{X}\|_F\right) \leq \epsilon_2$.
        $t \leftarrow t + 1$.
    end while
    Update the warp's parameters:
    $\{\mathbf{p}_i^c\}_{i=1}^N \leftarrow \{\mathbf{p}_i^c\}_{i=1}^N + \{\delta\mathbf{p}_i\}_{i=1}^N$
end while
```

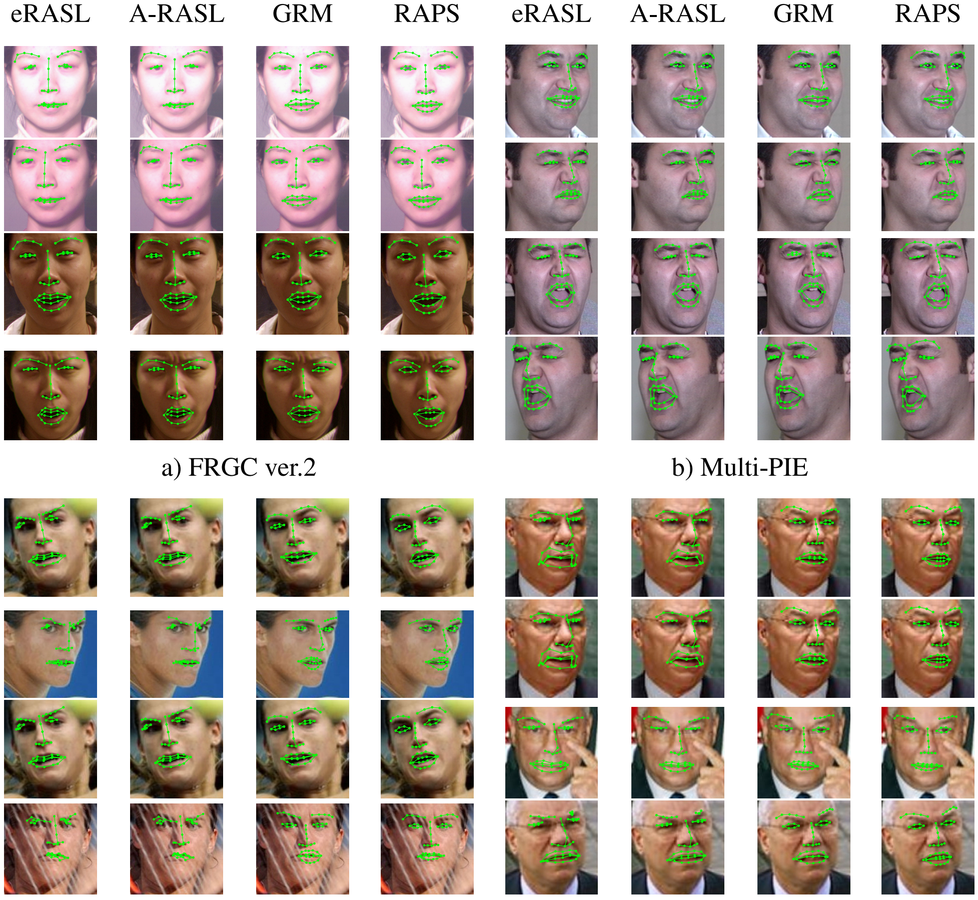
3.1. Image Congealing

The performance of the RAPS in the image congealing problem is assessed by conducting experiments on images taken from the Multi-PIE [14], the FRGC ver.2 [22], and the LFW [15] databases. The images of subject '002' ($N = 30$) which depict six different expressions under poses varying from $-30°$ to $30°$ were selected from Multi-PIE, while the available images ($N = 30$) of subject '04202' were selected from FRGC ver.2. The subjects 'Collins Powell' ($N = 34$) and 'Amelie Mauresmo' ($N = 21$) of the LFW databases were used for in-the-wild experiment. The ground-truth provided in [24] were used in order to evaluate the performance for the Multi-PIE and the FRGC ver.2 databases, respectively. In case of the LFW, the images were manually annotated with regards to 68 landmark points.

Given N input facial images of same subject, the initial estimation of the 68 landmark points position for each image of $\mathbf{X}$ were produced by the detector described in [31]. Subsequently, the basis $\mathbf{U}$ and the initial landmarks were

Table 1. Mean point-to-point error of 51 landmark points.

Subject	Initial	eRASL	A-RASL	GMR	RAPS	PSM/eRASL	PSM/A-RASL	PSM/GMR	PSM/RAPS
‘C. P.’	0.076	0.116	0.083	0.068	**0.054**	0.105	0.088	0.063	**0.045**
‘A.M.’	0.079	0.106	0.097	0.071	**0.056**	0.105	0.095	0.059	**0.053**
‘04202’	0.042	0.046	0.042	0.037	**0.034**	0.088	0.043	0.035	**0.032**
‘002’	0.048	0.041	0.041	0.042	**0.037**	0.043	0.042	0.041	**0.037**



a) FRGC ver.2

b) Multi-PIE

c) LFW

Figure 2. Sample fitting results from the Multi-PIE, the FRGC ver.2, and the LFW databases. The proposed method outperforms the compared methods [8,21,30] on images captured under challenging conditions such as expressions, illuminations, poses and occlusions.

given as input into Algorithm 1. Furthermore, the same initial landmarks were given as input in all the other tested methods, namely the eRASL, the A-RASL, and the GMR. The average point-to-point Euclidean distance of the 51 landmark points normalized by the Euclidean distance of the outer corner of eyes, was used as the error measure. The average errors for all tested methods are summarized in columns $2-5$ of Table 1, while the first column shows the initial error. By inspecting Table 1, the RAPS outperforms all the other methods by a large margin. More specifically, the RAPS achieved an average 25.1% accuracy improvement, while the GMR method achieved 10.9% improvement. This improvement is mainly due to the following two reasons. First, the GMR minimizes the non-regularized rank of the image ensemble which has been shown tends to unnaturally deform the subject’s facial appearance resulting in false face alignment [9]. Second, we explicitly include an error term that accounts for non-Gaussian errors/outliers robustifying the RAPS. Example images with the corresponding final position of landmark as produced by Algorithm 1, are depicted in Fig. 2. Finally, the convergence time (in CPU seconds) for all the methods is shown in Table 2. Clearly, RAPS outperforms all methods in terms of landmark localization accuracy, having also smaller complexity, since $N^2k \ll N^2F$.

Table 2. Overall computational time (in CPU seconds) for convergence.

Subject	RASL	A-RASL	GMR	RAPS
C. Powel	1078	444	350	**150**
A. Mauresmo	255	190	185	**168**
04202	203	135	170	**152**
002	235	196	210	**124**

3.2. Building Person Specific Models

In this section, the ability to build a PSM by using the results of batch alignment is investigated. Two different experiments were conducted by employing 1) still images and 2) video sequence, of the same subject.

Still images: Given the N images of the same subject, the problem of image congealing is solved by applying the RAPS. Furthermore, a leave-one-image-out experiment was performed using the corrected fittings. In particular, a PSM was built by excluding one of images at a time and use the excluded image as a test one. The same procedure was applied in all the compared methods. The average errors for all tested image sets are summarized in columns $6-9$ of Table 1. Clearly, the proposed method has a consistently better performance in terms of alignment accuracy compared to that obtained by the methods that is compared to. This also holds in case where the initialization performance was poor.

Video sequence: The ability of PSM trained from the results of RAPS is assessed by conducting facial features tracking experiments in the YouTube Celebrities Face Tracking and Recognition Dataset [16]. This dataset was collected from the internet and contains video sequences of celebrities captured under different in-the-wild conditions. Due to the fact that it was released as a dataset for face tracking and recognition the annotation of landmark points are not provided. To produce both quantitative and qualitative results 4[1] different sequences were annotated with regards to 51-landmarks.

Firstly, the initial position of landmark points for each frame of the video sequence was produced by [31]. Subsequently, the first 10% of the total number of frames were given into tested methods and the corrected shapes were used to build a PSM based on the output of each method. Finally, the residue frames were tracked by the PSMs and the registration errors of each frame were computed. Figure 3 plots the normalized point-to-point Euclidean error of 51 landmark points for all tested methods for each frame of the tested video sequences. By inspecting Fig. 3, it is clear that the proposed method outperforms the competing methods on the 4 videos. More specifically, the RAPS/PSM achieved 0.0404, 0.0421, 0.0624, 0.0598 mean error for each video sequence, while the GMR achieved 0.0481, 0.0511, 0.0719, 0.070, respectively. Visual tracking results produced by the PSM/RAPS are presented in Fig. 4. As it can be seen, the automatically constructed PSM produces consistent registration performance on video sequences with challenging variations such as poses, illumination conditions, and expressions.

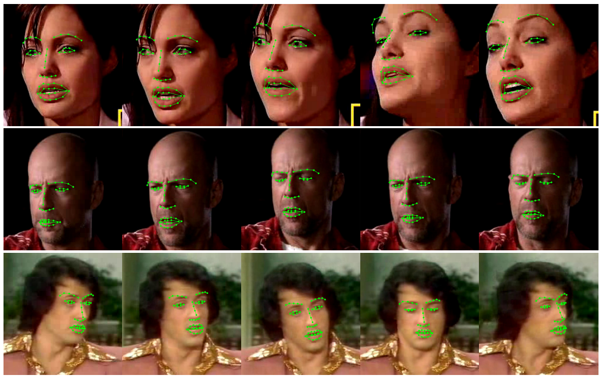


Figure 4. Example tracking results on the Youtube Celebrity Dataset.

4. Conclusions

In this paper, a robust and efficient method for the automatic construction of PSMs from erroneous initializations has been proposed. We show that, it is possible to use an orthonormal statistical prior of facial images to perform robust image congealing. This statistical prior prevents the unnatural deformations. We demonstrated that, the proposed method outperforms the state-of-the-art methods both in terms of computational efficiency and alignment accuracy. Furthermore, the potential of the method in automatically building a robust PSM which can be used for facial features tracking under unconstrained conditions, has been revealed.

Acknowledgements

The work of C. Sagonas and S. Zafeiriou was partially funded by the EPSRC project EP/J017787/1 (4DFAB). The work of Y. Panagakis was funded by the European Research Council under the FP7 Marie Curie Intra-European Fellowship. The work of M. Pantic was partially funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG).

References

[1] B. Amberg, A. Blake, and T. Vetter. On compositional image alignment, with an application to active appearance models. In *CVPR*, pages 1714–1721, 2009. 3

[1] 1) 0292_02_002_angelina_jolie, 2) 0502_01_005_bruce_willis, 3) 1621_02_017_ronald_reagan, and 4) 1786_02_006_sylvester_stallone

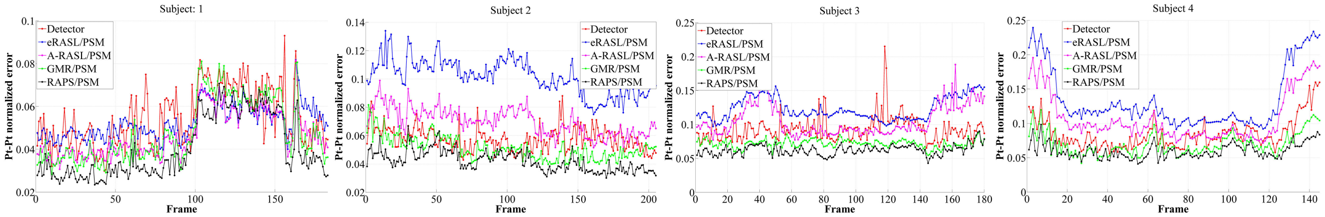


Figure 3. Average normalized point-to-point error produced by all tested methods for four video sequences of Youtube Celebrity Dataset.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pages 3444–3451, 2013. 1, 2

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pages 545–552, 2011. 1, 2

[4] D. P. Bertsekas. Constrained optimization and lagrange multiplier methods. *Computer Science and Applied Mathematics, Boston: Academic Press*, 1, 1982. 4

[5] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. 4

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011. 4

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, pages 2887–2894, 2012. 1

[8] X. Cheng, C. B. Fookes, S. Sridharan, J. Saragih, and S. Lucey. Deformable face ensemble alignment with robust grouped-l1 anchors. In *FG*, 2013. 2, 3, 5, 6

[9] X. Cheng, S. Sridharan, J. Saraghi, and S. Lucey. Anchored deformable face ensemble alignment. In *ECCV*, pages 133–142, 2012. 2, 3, 6

[10] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *IEEE TSMCB*, 42(4):1006–1016, 2012. 2

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, 2001. 1, 2, 3

[12] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995. 1

[13] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 17, pages 929–938, 2006. 1

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *IMAVIS*, 28(5):807–813, 2010. 5

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, October 2007. 5

[16] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, pages 1–8, 2008. 7

[17] A. Levy and M. Lindenbaum. Sequential karhunen-loeve basis extraction and its application to images. In *ICIP*, volume 2, pages 456–460, 1998. 2

[18] X. Liu. Discriminative face alignment. *IEEE TPAMI*, 31(11):1941–1954, 2009. 2

[19] S. Lucey, S. Sridharan, R. Navarathna, and A. B. Ashraf. Fourier lucas-kanade algorithm. *IEEE TPAMI*, 35(6):1383–1396, 2013. 1, 2

[20] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004. 1, 2, 3

[21] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE TPAMI*, 34(11):2233–2246, 2012. 2, 3, 5, 6

[22] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR*, volume 1, pages 947–954, 2005. 5

[23] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV-W*, 2013. 5

[24] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR-W*, pages 896–903, 2013. 5

[25] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011. 1, 2

[26] J. Sung and D. Kim. Adaptive active appearance model with incremental learning. *Pattern Recognition Letters*, 30(4):359–367, 2009. 2

[27] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, and M. Pantic. Generic active appearance models revisited. In *ACCV*, pages 650–663. 2012. 1, 2

[28] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. *ACM Trans. on Graphics*, 30(4):77, 2011. 2

[29] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 1, 2

[30] C. Zhao, W.-K. Cham, and X. Wang. Joint face alignment with a generic deformable face model. In *CVPR*, pages 561–568, 2011. 2, 3, 4, 5, 6

[31] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012. 5, 7