Bayesian Active Appearance Models

Joan Alabort-i-Medina Stefanos Zafeiriou
Department of Computing, Imperial College London, United Kingdom
{ja310,s.zafeiriou}@imperial.ac.uk

Abstract

In this paper we provide the first, to the best of our knowledge, Bayesian formulation of one of the most successful and well-studied statistical models of shape and texture, i.e. Active Appearance Models (AAMs). To this end, we use a simple probabilistic model for texture generation assuming both Gaussian noise and a Gaussian prior over a latent texture space. We retrieve the shape parameters by formulating a novel cost function obtained by marginalizing out the latent texture space. This results in a fast implementation when compared to other simultaneous algorithms for fitting AAMs, mainly due to the removal of the calculation of texture parameters. We demonstrate that, contrary to what is believed regarding the performance of AAMs in generic fitting scenarios, optimization of the proposed cost function produces results that outperform discriminatively trained state-of-the-art methods in the problem of facial alignment "in the wild".

1. Introduction

The construction and fitting of deformable models is a very active area of research in computer vision because of its great importance in robust articulated object detection, recognition and tracking. One of the most well-studied technique for building and fitting deformable models are Active Appearance Models (AAMs) [1, 2] and the closely related 3D Morphable Models [3]. AAMs use statistical models to describe shape and texture variation. In particular, a statistical model of shape is built from a set of (manually) annotated fiducial points describing the shape of the object of interest. In order to approximately retain only the variability that is attributed to non-rigid deformations, the shape points are put in correspondence (usually by removing global similarity transforms using Generalized Procustes Analisys [2]) [1].

Similarly, a statistical model of the texture is built using images of the object that have been normalized with respect to the shape points (so-called shape-free textures). This requires a predefined reference frame (usually defined in terms of the mean shape) and a global motion model or warp $\mathcal{W}(\mathbf{p})$ (e.g. Piece-Wise Affine [2] or Thin Plate Spline [1, 4]). The two main assumptions behind AAMs are that (1) for every test (unseen) image there exists a test shape and set of texture weights for which the test shape can be warped onto the reference frame and expressed as a linear combination of the shape-free training textures and (2) the test shape can be written as a linear combination of the training shapes. In mathematical terms let $\mathcal{S} = \{\bar{\mathbf{s}}, \mathbf{B} \in \Re^{2p \times n}\}$ and $\mathcal{T} = \{\bar{\mathbf{m}}, \mathbf{U} \in \Re^{F \times m}\}$ be the linear models for the shape and texture, respectively (where p is the number of shape points, F the number of pixels on the reference frame and n and m denote the number of bases of the shape and texture models, respectively). Then, according to the above assumptions, given a test shape $\mathbf{s} \in \Re^{2p \times 1}$ and its corresponding test image $\mathbf{x}$ we have the two following approximation

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

where $\mathbf{x}(\mathcal{W}(\mathbf{p})) \in \Re^{F \times 1}$ is the vectorized shape-free texture of the test image (from now onwards, for simplicity, we will write $\mathbf{x}(\mathbf{p})$ instead of $\mathbf{x}(\mathcal{W}(\mathbf{p}))$). Under the previous assumption the parameters $\mathbf{p}$ and $\mathbf{c}$ are retrieved by

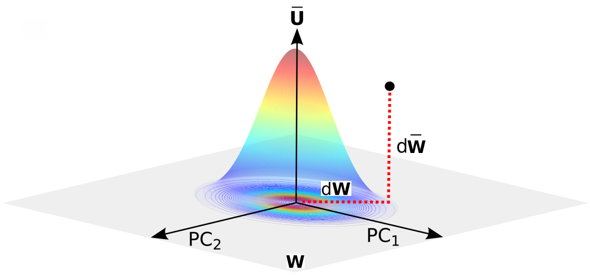

Figure 1: Our Bayesian formulation fits AAMs by minimizing two different distances: (i) the Mahalanobis distance within the latent texture subspace $\mathbf{W}$ and (ii) the Euclidean distance within its orthogonal complement $\bar{\mathbf{W}}$ weighted by the inverse of the estimated sample noise.

[1] Dense shape models such as 3D Morphable Models use more complicated procedures to arrange the shapes in correspondence [3].

minimizing the sum of squared errors between the previous shape-free texture and its reconstruction by the statistical texture model

$$\mathbf{p}_o, \mathbf{c}_o = \arg\min_{\mathbf{p},\mathbf{c}} ||\mathbf{x}(\mathbf{p}) - (\mathbf{m} + \mathbf{U}\mathbf{c})||^2_{\mathbf{P}} \qquad (2)$$

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.

Several works have been proposed to solve the previous optimization problem [5, 4, 2, 1]. Most notable methodologies include the regression-based method of [1], the very fast project-out inverse compositional algorithm (PIC) [2] (which has been heavily criticized for its inability to perform well under generic fitting scenarios, i.e., fit images of unseen identities), the simultaneous inverse compositional algorithm [5], and a variation of the simultaneous inverse compositional algorithm that operates in a projected space [4]. A complete project-out compositional framework for fitting AAMs was proposed in [6]. Due to the popularity of the PIC algorithm [2], mainly because of its extremely low computational complexity, methodologies such as [4, 6], which can provide near real-time fitting, have not received much attention.

AAMs are often criticized for a variety of reasons. The most common is that defining a linear statistical model of texture that explains variations in identity, expressions, pose and illumination, is a very challenging task, especially in the intensity domain. Furthermore, the large variation in facial appearance makes it very difficult to perform regression from texture differences to shape parameters. Additionally, occlusion cannot be easily handled, and, in general, require the application of robust estimators on the ℓ_2^2-loss function in Eq. (2). Finally, joint optimization with respect to shape and texture parameters may create numerous local minima in the cost function making it difficult for the algorithms to reach optimal solutions.

Due to the above limitations, recent research on facial alignment has focused on generative and discriminant fitting of part-based models [7, 8, 9] (*i.e.* models which do not define a complete holistic texture model of the object) and on regression-based techniques that directly learn mappings from image features to shape parameters or landmark locations [10, 11]. The main advantages of part-based models are a natural handling of partial occlusions (since they only model certain parts of the object) and, most importantly, the fact that they are optimized only with respect to shape (they do not define parametric models of texture). Notable examples include Constrained Local Models (CLMs) [7] and the tree-based model of [8] (which can be also used for object detection). More recently, Asthana *et al.* [9] proposed a robust discriminative framework for fitting CLMs which achieved state-of-the-art results in the problem of facial alignment “in the wild”. On the other hand, recent regression-based approaches have focused on combining cascade-regression methods with the use of highly engineered nonlinear image features [10, 11]. In particular, the supervised descent method of [11], which learns a sequence of simple linear regressors from SIFT features to a non-parametric shape representation, is considered to be the state-of-the-art approach to facial alignment “in the wild”.

In this paper, we examine the problem of fitting AAMs under a Bayesian perspective. Summarizing, our key contributions are:

- To provide a novel Bayesian formulation of AAMs. To this end, we use a simple probabilistic model for texture generation assuming both Gaussian noise and a Gaussian prior over a latent texture space (i.e., $\mathbf{c}$). By marginalizing out the latent texture space, we derive a novel cost function that only depends on shape parameters and propose an efficient compositional algorithm to optimize it (the proposed cost function is motivated by seminal works on probabilistic component analysis and object tracking [12, 13, 14])

- To present the first in-depth comparison between the existent gradient descent algorithms for fitting AAMs [2, 5, 4, 6] to images acquired “in the wild”.

- To show that our Bayesian AAM approach (and others [4]) can outperform state-of-the-art methods in facial alignment such as the Robust Discriminative Response Map Fitting (DRMF) for CLMs of [9] and the Supervised Descent Method (SDM) of [11].

The remainder of the paper is structured as follows. Section 2 reviews Principal Component Analysis (PCA) and Probabilistic PCA (PPCA). Section 3 outlines the existent gradient descent algorithms for fitting AAMs. Our novel Bayesian AAM formulation is introduced in Section 4. Experimental results are shown in Section 5. Finally, conclusions are drawn in Section 6.

2. Principal Component Analysis (PCA) and Probabilistic PCA

In the majority of cases, the statistical models of shape and texture used in AAMs are defined using Principal Component Analysis [15]. In this section we will briefly review both deterministic and probabilistic versions of PCA.

The deterministic version of PCA finds a set of orthonormal projection bases $\mathbf{U}$ so that the latent space $\mathbf{C}$ is the projection of the mean-centered training set $\bar{\mathbf{X}} = [\mathbf{x}_1 - \mathbf{m}, \ldots, \mathbf{x}_n - \mathbf{m}]$ onto $\mathbf{U}$ (*i.e.* $\mathbf{C} = \mathbf{U}^T\bar{\mathbf{X}}$). The optimization problem is defined as follows

$$\mathbf{U}_o = \arg\max_{\mathbf{U}} \operatorname{tr}\left[\mathbf{U}^T\mathbf{S}\mathbf{U}\right], \text{ s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I} \qquad (3)$$

where $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{T}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T = \frac{1}{N}\bar{\mathbf{X}}\bar{\mathbf{X}}$ is the total scatter matrix and $\mathbf{I}$ the identity matrix. The optimal M projection bases matrix $\mathbf{U}_o$ are recovered by keeping the M eigenvectors $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]$ that correspond to the M largest eigenvalues of $\mathbf{S}$ (in the following we will assume that the eigenvalues are stored in a diagonal matrix $\mathbf{\Lambda} = \text{diag}\{[\lambda_1, \ldots, \lambda_M]\}$).

Probabilistic versions of PCA (PPCA) were independently proposed in [14, 12, 13] [2]. In these works the following probabilistic generative model was defined:

$$\begin{aligned} \mathbf{x} &= \mathbf{W}\mathbf{c} + \mathbf{m} + \boldsymbol{\epsilon} \\ \mathbf{c} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \end{aligned} \tag{4}$$

where $\mathbf{W}$ is the matrix that relates the latent variables $\mathbf{c}$ with the observed sample $\mathbf{x}$ and $\boldsymbol{\epsilon}$ is the sample noise which is assumed to be an isotropic Gaussian. The motivation is that, when $N < F$, the latent variables will offer a more compact representation of the dependencies between the observations. Denoting the parameters as $\theta = \{\mathbf{W}, \sigma^2, \mathbf{m}\}$, the posterior probability over the latent variables is given by

$$p(\mathbf{c}|\mathbf{x}, \theta) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \mathbf{m}), \sigma^2\mathbf{M}^{-1}) \tag{5}$$

where $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$ (note that here the bases $\mathbf{W}$ are not required to be orthonormal). Using a Maximum Likelihood (ML) approach the parameters θ are found by solving

$$\begin{aligned} \theta_o &= \arg\min_{\theta}\ \ln \textstyle\prod_{i=1}^{n} p(\mathbf{x}_i|\theta) \\ &= \arg\min_{\theta}\ \ln \textstyle\prod_{i=1}^{n} \int_{\mathbf{c}} p(\mathbf{x}_i|\mathbf{c}, \theta)\, p(\mathbf{c})\, d\mathbf{c}. \end{aligned} \tag{6}$$

with the optimal $\mathbf{W}$, σ^2 and $\mathbf{m}$ given by

$$\begin{aligned} \mathbf{W} &= \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{1/2}\mathbf{R} \\ \sigma^2 &= \tfrac{1}{N-M}\textstyle\sum_{j=M+1}^{N}\lambda_i \\ \mathbf{m} &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{x}_i. \end{aligned} \tag{7}$$

where $\mathbf{R}$ is an arbitrary $M \times M$ orthogonal matrix.

An alternative to the ML approach is to use an Expectation Maximization (EM) procedure where the first and second order moments of the latent space ($\mathbb{E}[\mathbf{c}_i]$ and $\mathbb{E}[\mathbf{c}_i\mathbf{c}_i^T]$) are also found. The EM solutions for the parameters can be found in [16, 12]. Several variations of probabilistic PCA have been proposed, e.g. by incorporating sparseness and nonnegative constraints [17] or changing the Gaussian models for others (such as the Student-t model of [18]).

In the upcoming sections, we first review existent algorithms for fitting AAMs and then proceed to define a novel fitting strategy using the previous probabilistic model as a generative model of texture.

3. Fitting Active Appearance Models

Before we describe our Bayesian AAM formulation, we briefly outline the main algorithms that have been proposed to solve the optimization problem defined by Eq. (2). In this paper, we limit the discussion to compositional Gauss-Newton algorithms [2, 5, 4, 6] for fitting AAMs and, consequently, we will not review discriminative and regression based approaches. For such methods the reader is referred to the existent literature [19, 1, 20, 21].

In a compositional framework, Gauss-Newton algorithms iteratively solve Eq. (2) with respect to the shape parameters $\mathbf{p}$ by (i) introducing an incremental warp $\mathcal{W}(\delta\mathbf{p})$, (ii) performing a first order Taylor expansion of the residual term with respect to $\delta\mathbf{p}$, (iii) solving for $\delta\mathbf{p}$ and (iv) computing the optimal shape parameters $\mathbf{p}_o$ by composing the incremental warp $\mathcal{W}(\delta\mathbf{p}_o)$ with the current estimate of the warp $\mathcal{W}(\mathbf{p}_c)$. Depending on whether the algorithm is forward or inverse the incremental warp is placed on the image (forward) or model (inverse) side, and the corresponding linearization and composition performed according to this choice. In both settings, the incremental warp is linearized around the identity warp (denoted by $\delta\mathbf{p} = \mathbf{0}$).

Mathematically, the forward setting is defined by the following linearizations and update rules

$$\begin{aligned} \mathbf{x}(\mathbf{p}_c \circ \delta\mathbf{p}) - (\mathbf{m} + \mathbf{U}\mathbf{c}) \;&\approx\; \mathbf{e}(\mathbf{p}_c) \;+ \mathbf{J}_{\mathbf{x}}\delta\mathbf{p} \\ \mathcal{W}(\mathbf{p}_o) \;&\leftarrow\; \mathcal{W}(\mathbf{p}_c) \circ \mathcal{W}(\delta\mathbf{p}_o) \end{aligned} \tag{8}$$

and the inverse

$$\begin{aligned} \mathbf{x}(\mathbf{p}_c) - (\mathbf{m}(\delta\mathbf{p}) + \mathbf{U}(\delta\mathbf{p})\mathbf{c}) \;&\approx\; \mathbf{e}(\mathbf{p}_c) \;- \mathbf{J}_{\mathbf{m}}\delta\mathbf{p} \\ \mathcal{W}(\mathbf{p}_o) \;&\leftarrow\; \mathcal{W}(\mathbf{p}_c) \circ \mathcal{W}(\delta\mathbf{p}_o)^{-1} \end{aligned} \tag{9}$$

where $\mathbf{e}(\mathbf{p}_c) = \mathbf{x}(\mathbf{p}_c) - (\mathbf{m} + \mathbf{U}\mathbf{c})$ is the so-called error image and where $\mathbf{J}_{\mathbf{x}}, \in \Re^{F\times n}$ and $\mathbf{J}_{\mathbf{m}}, \in \Re^{F\times n}$ are the image and model Jacobians evaluated at $\delta\mathbf{p} = \mathbf{0}$, respectively. By using the chain rule, the previous Jacobians can be further expanded as $\nabla_W \mathbf{x}\frac{\partial W}{\partial \mathbf{p}}$ and $\nabla_W(\mathbf{m} + \mathbf{U}\mathbf{c})\frac{\partial W}{\partial \mathbf{p}}$. For further details on how to compute $\frac{\partial W}{\partial \mathbf{p}}$ and on warp composition and inversion the interested reader is referred to [2] and [4]. In general, the optimization of Eq. (2) with respect to the texture parameters $\mathbf{c}$ is algorithm dependent.

3.1. Project Out Inverse Compositional

The most popular and fastest algorithm for solving Eq. (2) is the so-called project out inverse compositional algorithm [3] which was first proposed in [22] for performing rigid alignment with linear texture variations. This algorithm eliminates the need to solve for the texture parameters $\mathbf{c}$ by working on the orthogonal complement of the texture subspace $\mathbf{U}$ (*i.e.* $\bar{\mathbf{U}} = \mathbf{I} - \mathbf{U}\mathbf{U}^T$). Consequently, the incremental warp $\delta\mathbf{p}$ is estimated only using the mean $\mathbf{m}$ of the

[2] In particular the Maximum Likelihood (ML) solutions where provided in [12, 13], while Expectation Maximization (EM) solutions where presented in [12, 14]

[3] Inverse compositional algorithms became very popular after [2].

texture model

$$\delta\mathbf{p}_o = \arg\min_{\delta\mathbf{p}} ||\mathbf{x}(\mathbf{p}_c) - \mathbf{m}(\delta\mathbf{p})||^2_{\mathbf{I}-\mathbf{U}\mathbf{U}^T} \tag{10}$$

The problem is solved by linearizing over $\mathbf{m}(\delta\mathbf{p}) = \mathbf{m} + \mathbf{J}_\mathbf{m}\delta\mathbf{p}$. By defining $\tilde{\mathbf{J}}_\mathbf{m} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{J}_\mathbf{m}$ the iterative updates are given by

$$\delta\mathbf{p} = (\tilde{\mathbf{J}}_\mathbf{m}^T\tilde{\mathbf{J}}_\mathbf{m})^{-1}\tilde{\mathbf{J}}_\mathbf{m}^T(\mathbf{x}(\mathbf{p}_c) - \mathbf{m}) \tag{11}$$

The computation of $(\tilde{\mathbf{J}}_\mathbf{m}^T\tilde{\mathbf{J}}_\mathbf{m})^{-1}\tilde{\mathbf{J}}_\mathbf{m}^T$ is performed offline, hence the complexity of each update is only $O(nN + n^2)$.

3.2. Simultaneous Inverse Compositional

The simultaneous inverse compositional algorithm [5], finds, simultaneously, shape and texture increments ($\delta\mathbf{p}$ and $\delta\mathbf{c}$) by solving the following optimization problem

$$\delta\mathbf{p}_o, \delta\mathbf{c}_o = \arg\min_{\delta\mathbf{p},\delta\mathbf{c}} ||\mathbf{x}(\mathbf{p}_c) - (\mathbf{m}(\delta\mathbf{p}) + \mathbf{U}(\delta\mathbf{p})(\mathbf{c}_c + \delta\mathbf{c})||^2 \tag{12}$$

Let $\delta\mathbf{q} = [\delta\mathbf{p}^T, \delta\mathbf{c}^T]^T$ be the concatenation of the parameters and let $\mathbf{U}(\delta\mathbf{p}) \approx \mathbf{U} + [\mathbf{J}_1\delta\mathbf{p} \ldots \mathbf{J}_m\delta\mathbf{p}]$ be the linearization of the bases, where $\mathbf{J}_i$ is the Jacobian with respect to each component $\mathbf{u}_i$. The updates of the parameters are given by

$$\delta\mathbf{q} = (\mathbf{J}_\mathbf{t}^T\mathbf{J}_\mathbf{t})^{-1}\mathbf{J}_\mathbf{t}^T(\mathbf{x}(\mathbf{p}_c) - (\mathbf{m} + \mathbf{U}\mathbf{c}) \tag{13}$$

where the total Jacobian $\mathbf{J}_\mathbf{t} = [\mathbf{J}_\mathbf{U}, \mathbf{U}]$ and $\mathbf{J}_\mathbf{U} = \mathbf{J}_\mathbf{m} + \sum_{i=1}^K \mathbf{c}_c^i\mathbf{J}_i$ (neglecting second order terms of the form $\delta\mathbf{c}^T\mathbf{A}\delta\mathbf{p}$). Even though $\mathbf{J}_\mathbf{m}$ and all individual Jacobians $\mathbf{J}_i$ can be precomputed, the computation of the $\mathbf{J}_\mathbf{t}$ must be performed at each step due to its dependency on the current estimate of the texture parameters $\mathbf{c}_c$. Thus, the total cost per iteration is $O((n+m)^2N + (n+m)^3)$

3.3. Alternating Optimization Approaches

The variation of the simultaneous inverse compositional algorithm proposed in [4] solves two different problems, in an alternating manner, one for the shape and one for the appearance, as

$$\begin{aligned}\delta\mathbf{p}_o &= \arg\min_{\delta\mathbf{p}} ||\mathbf{x}(\mathbf{p}_c) - (\mathbf{m}(\delta\mathbf{p}) + \mathbf{U}(\delta\mathbf{p})\mathbf{c}_c||^2_{\mathbf{I}-\mathbf{U}\mathbf{U}^T} \\ \delta\mathbf{c}_o &= \arg\min_{\delta\mathbf{c}} ||\mathbf{x}(\mathbf{p}_c) - (\mathbf{m}(\delta\mathbf{p}_o) + \mathbf{U}(\delta\mathbf{p}_o)(\mathbf{c}_c + \delta\mathbf{c}))||^2\end{aligned} \tag{14}$$

The update for $\delta\mathbf{p}$ is given by

$$\delta\mathbf{p} = (\tilde{\mathbf{J}}_\mathbf{U}^T\tilde{\mathbf{J}}_\mathbf{U})^{-1}\mathbf{J}_\mathbf{U}^T(\mathbf{x}(\mathbf{p}_c) - \mathbf{m}) \tag{15}$$

where $\tilde{\mathbf{J}}_\mathbf{U} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{J}_\mathbf{U}$. Similarly as before, $\tilde{\mathbf{J}}_\mathbf{m}$ and all $\tilde{\mathbf{J}}_i$ can be precomputed, but $\tilde{\mathbf{J}}_\mathbf{U}$ has to be computed at each iteration because of its dependancy on $\mathbf{c}_c$. Given the optimum $\delta\mathbf{p}_o$, $\delta\mathbf{c}_0$ is obtained by solving the second optimization problem in Eq. (14)

$$\delta\mathbf{c} = \mathbf{U}^T(\mathbf{x}(\mathbf{p}_c) - (\mathbf{m} + \mathbf{U}\mathbf{c}_c + \mathbf{J}_\mathbf{U}\delta\mathbf{p}_o)) \tag{16}$$

By performing expansions of the update in Eq. (15) with regards to the projection operation $\mathbf{I} - \mathbf{U}\mathbf{U}^T$, it has been shown [4] that the method is of complexity $O(m^2n^2 + (m+n)N + n^3)$.

4. Probabilistic Models for fitting AAMs

Let us consider again the probabilistic generative model defined in Eq. 4. For a particular test image $\mathbf{x}(\mathbf{p})$ we have

$$\begin{aligned}\mathbf{x}(\mathbf{p}) &= \mathbf{W}\mathbf{c} + \mathbf{m} + \boldsymbol{\epsilon} \\ \mathbf{c} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})\end{aligned} \tag{17}$$

where $\mathbf{W}, \mathbf{m}$ and σ^2 have been learned in the training phase (from a set of training shape-free textures). Notice that we have changed the prior over the latent space such that it is a multivariate Gaussian distribution with variance equal to the eigenvalues $\boldsymbol{\Lambda}$ (instead of $\mathbf{I}$). Furthermore, for simplicity and without loss of generality, we assume that $\mathbf{W}$ is orthonormal (i.e., $\mathbf{W}^T\mathbf{W} = \mathbf{I}$).

Our aim is to define a ML procedure to retrieve the optimal shape parameters $\mathbf{p}_o$ using the above generative model. This can be done by defining the following optimization problem

$$\begin{aligned}\mathbf{p}_o &= \arg\max_{\mathbf{p}} \ \ln p(\mathbf{x}(\mathbf{p})|\theta) \\ &= \arg\max_{\mathbf{p}} \ \ln \int_\mathbf{c} p(\mathbf{x}(\mathbf{p})|\mathbf{c}, \theta)p(\mathbf{c}|\theta)d\mathbf{c}\end{aligned} \tag{18}$$

where the texture parameters $\mathbf{c}$ are marginalized out and the marginalized density $p(\mathbf{x}(\mathbf{p})|\theta)$ is given by

$$p(\mathbf{x}(\mathbf{p})) = \mathcal{N}(\mathbf{m}, \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^T + \sigma^2\mathbf{I}) \tag{19}$$

Using Eq. (19), the optimization problem in Eq.(18) can be reformulated as

$$\begin{aligned}\mathbf{p}_o &= \arg\min_{\mathbf{p}} \ ||\mathbf{x}(\mathbf{p}) - \mathbf{m}||^2_{(\mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}} \\ &= \arg\min_{\mathbf{p}} \ ||\mathbf{x}(\mathbf{p}) - \mathbf{m}||^2_{\mathbf{W}\mathbf{D}^{-1}\mathbf{W}^T} + \\ &\qquad \tfrac{1}{\sigma^2}||\mathbf{x}(\mathbf{p}) - \mathbf{m}||^2_{\mathbf{I}-\mathbf{W}\mathbf{W}^T}\end{aligned} \tag{20}$$

where we used the Woodbury formula

$$\begin{aligned}(\mathbf{W}\mathbf{L}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1} = \mathbf{W}\mathbf{D}^{-1}\mathbf{W}^T + \\ \tfrac{1}{\sigma^2}(\mathbf{I} - \mathbf{W}\mathbf{W}^T)\end{aligned} \tag{21}$$

where $\mathbf{D}$ is a diagonal matrix defined as $\mathbf{D} = \mathrm{diag}[\lambda_1 + \sigma^2, \cdots, \lambda_M + \sigma^2]$.

Hence, as in [13], our cost function is comprised of two different distances: (i) the Mahalanobis distance within the latent texture subspace $\mathbf{W}$ and (ii) the Euclidean distance within its orthogonal complement $\bar{\mathbf{W}} = \mathbf{I} - \mathbf{W}\mathbf{W}^T$ weighted by the inverse of the estimated sample noise σ^2. The first of these distances favors solutions with higher probability within latent subspace $\mathbf{W}$, acting as a regularizer that ensures the solution $\mathbf{x}(\mathbf{p}_o)$ can be well reconstructed by the texture model. The second distance captures everything that cannot be generated by the texture model (*e.g.* occlusions and other unseen variations) and weights it with respect to the estimated sample noise [4].

Note that, the contribution of the second term $\frac{1}{\sigma^2}||\mathbf{x}(\mathbf{p}) - \mathbf{m}||^2_{\mathbf{I}-\mathbf{W}\mathbf{W}^T}$ decreases as the estimated sample noise increases. On the other hand, when the variance $\mathbf{\Lambda}$ of the prior over the latent subspace increases (and especially as $\mathbf{\Lambda} \rightarrow \infty$) $\mathbf{c}$ becomes uniformly distributed and the contribution of the first term $||\mathbf{x}(\mathbf{p}) - \mathbf{m}||^2_{\mathbf{W}\mathbf{D}^{-1}\mathbf{W}^T}$ vanishes. Hence, under our Bayesian formulation, the project-out inverse compositional algorithm in Section 3.1 naturally stems from assuming a uniform prior over the latent texture space.

On the contrary, our ML formulation uses both distances to retrieve the optimal shape parameters. To the best of our knowledge the above cost function has not been used for estimating parameters in a deformable model fitting framework.

A novel forward compositional algorithm is proposed to solve the optimization problem defined in Eq. (20). As we demonstrate below, the algorithm is efficient as only shape parameters need to be recovered

$$\delta\mathbf{p}_o = \arg\min_{\delta\mathbf{p}} \; ||\mathbf{x}(\mathbf{p}_c \circ \delta\mathbf{p}) - \mathbf{m}||^2_{\mathbf{W}\mathbf{D}^{-1}\mathbf{W}^T} + \frac{1}{\sigma^2}||\mathbf{x}(\mathbf{p}_c \circ \delta\mathbf{p}) - \mathbf{m}||^2_{\mathbf{I}-\mathbf{W}\mathbf{W}^T} \qquad (22)$$

The problem can be solved by linearizing over $\mathbf{x}(\mathbf{p}_c \circ \delta\mathbf{p}) \approx \mathbf{x}(\mathbf{p}_c) + \mathbf{J}_\mathbf{x}\delta\mathbf{p}$ and the update is given by

$$\delta\mathbf{p} = (\tilde{\mathbf{J}}_\mathbf{x}^T\tilde{\mathbf{J}}_\mathbf{x})^{-1}\tilde{\mathbf{J}}_\mathbf{x}^T(\mathbf{x}(\mathbf{p}_c) - \mathbf{m}) \qquad (23)$$

where $\tilde{\mathbf{J}}_\mathbf{x} = (\mathbf{W}\mathbf{D}^{-1}\mathbf{W}^T + \frac{1}{\sigma^2}(\mathbf{I} - \mathbf{W}\mathbf{W}^T))\mathbf{J}_\mathbf{x}$. In order to reduce the computation complexity of the algorithm we expand the Hessian as

$$\tilde{\mathbf{J}}_\mathbf{x}^T\tilde{\mathbf{J}}_\mathbf{x} = \frac{1}{\sigma^2}\mathbf{J}_\mathbf{x}^T\mathbf{J}_\mathbf{x} - (\mathbf{W}^T\mathbf{J}_\mathbf{x})^T\mathbf{A}\mathbf{W}^T\mathbf{J}_\mathbf{x} \qquad (24)$$

where $\mathbf{A} = (\frac{1}{\sigma^2}\mathbf{I} - \mathbf{D}^{-1})$. Note that the term $\tilde{\mathbf{J}}_\mathbf{x}^T(\mathbf{x}(\mathbf{p}_c) - \mathbf{m})$ can also be reformulated as $\tilde{\mathbf{J}}_\mathbf{x}^T(\mathbf{x}(\mathbf{p}_c) - \mathbf{m}) = \mathbf{J}_\mathbf{x}^T\tilde{\mathbf{x}}(\mathbf{p}_c)$ where

$$\tilde{\mathbf{x}}(\mathbf{p}_c) = \frac{1}{\sigma^2}(\mathbf{x}(\mathbf{p}_c) - \mathbf{m}) - (\mathbf{W}^T(\mathbf{x}(\mathbf{p}_c) - \mathbf{m}))^T\mathbf{A}\mathbf{W}^T(\mathbf{x}(\mathbf{p}_c) - \mathbf{m}) \qquad (25)$$

Using the above expansions the computational complexity of the update is of order $O(mnK + n^2K + n^3)$, which is similar or even lower (as m increases) to the complexity of [4].

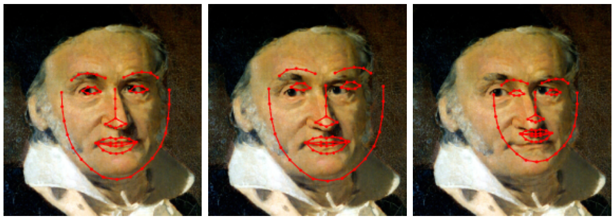
$\sigma_{\text{noise}} = 0$ $\sigma_{\text{noise}} = 4$ $\sigma_{\text{noise}} = 10$

Figure 2: Exemplar initializations obtained by varying the value σ_{noise} in the described initialization procedure. Note that, when $\sigma_{\text{noise}} = 0$ the initialization is equivalent to applying the correct scale and translation transforms to the mean shape. On the other hand, increasing values of σ_{noise} produce more challenging initialization.

We would like to clarify that our approach susbtancially differs from previous probabilistic formulations of AAMs [24]. In [24] a methodology that uses a K-mixture of PPCA to define AAMs is proposed. This mixture-AAM is fitted to new images by independently fitting K different AAMs (one for each mixture) using the project-out inverse compositional algorithm (PIC). In a way, [24] can be considered to automatize previous work in view-based AAMs where different AAMs (mainly view specific [25]) are independently fitted to a novel image. Consequently, in the one mixture case, the algorithm is equivalent to PIC. Finally, we want to note that multi-view and mixture of subspaces can be also used in our Bayesian framework, but we opted to show the power of our approach in the most difficult case, i.e. by using a single subspace (which is also faster and easier to build and fit).

5. Experiments

In this section we evaluate the performance of our Bayesian AAMs formulation on the problem of facial alignment "in the wild".

We performed two different experiments. The first one compares the proposed methodology with other existent gradient descent algorithms for fitting AAMs [2, 5, 4] on the popular LFPW [26] database. In the second experiment, we test our approach against two recently proposed state-of-art methods [9, 11] for facial alignment by performing two challenging cross-database experiments on the recently proposed Helen [27] and AFW [8] databases. Performance in both experiments is reported using the error measure proposed in [8] for the 49 interior points (excluding the face boundary) shown in figures Fig. 2 and Fig 7.

[4] This weighting corrects the size of the Gauss-Newton step taken within the orthogonal subspace $\bar{\mathbf{W}}$ (effectively addressing a well known problem of the original inverse compositional algorithm [23])

5.1. Comparison with other AAM fitting algorithms

We start by evaluating the relative performance of the proposed algorithm with respect to the existent Gradient Descent algorithms reviewed in Section 3 (*i.e.* the project-out inverse compositional algorithm [2], the simultaneous inverse compositional algorithm [5] and the alternative simultaneous inverse compositional algorithm proposed in [4] which we abbreviate as PIC, SIC and AIC respectively; the proposed algorithm is abbreviated as PROB).

This experiment is performed on the popular Labeled Faces Parts in the Wild (LFPW) [26] database. The original LFPW database consisted of 1400 URLs to images, 1100 for training and 300 for testing, that could be downloaded from the Internet. All images were acquired "in the wild" and contain large variations in identity, pose, illumination, expression and occlusion. Unfortunately some of the original URLs are no longer valid. We were able to download 813 training images and 224 test images for which we used the 68-point annotations provided by the authors of [28, 29] (which can be downloaded from [30]). All methods were trained using the available 813 training images and results are reported on the 224 testing images that remain available. For this experiment, we used Normalized pixel Intensities (NI) as the texture representation used to build the texture model of all AAM algorithms. Furthermore, in order to provide a little insight on the convergence properties of our method, given a particular test image, all methods are initialized by randomly perturbing the correct global similarity transform (without considering in-plane rotations) and applying it to the mean shape of the shape model (a frontal pose and neutral expression looking shape). The similarity transform is perturbed using a similar procedure as the one described in [2], where the parameter σ_{noise} controls the magnitude of the random Gaussian noise added to perform the perturbation. Exemplar initializations obtained by this procedure are shown in Fig. 2.

Results for this experiment are shown in Fig. 3 and Fig. 4 (for visual inspection, please see the fitting results in Fig. 7 and our supplementary material). Fig. 3 shows the Cumulative Error Distribution (CED) curves obtained by initializing all methods using $\sigma_{\text{noise}} = 4$. Fig. 4 shows the error bars and median error of the SIC-NI, AIC-NI and PROB-IC methods (PIC-NI did not fit on the graph) for increasing values of σ_{noise}. The results show that our approach (PROB-NI) considerably outperforms all other methods by a considerable large margin. More specifically, our Bayesian algorithm achieves improvements of at least 10% over all other algorithms at the significant region $0.025 < err > 0.035$ (region at which results are generally considered adequate by visual inspection). It is also worth noticing the good performance achieved by the alternating inverse compositional algorithm (AIC-NI) which, to our surprise, had barely been used in the AAMs literature before. Notice that, both PROB-NI and AIC-NI are also fairly robust against the magnitude of σ_{noise}. Finally, the remaining AAMs algorithms: the simultaneous inverse compositional (SIC-NI) and specially the project-out inverse compositional (PIC-NI) perform poorly.

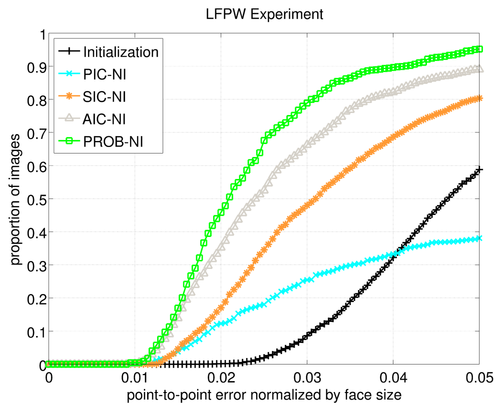

Figure 3: CED curves for $\sigma_{\text{noise}} = 4$ on the LFPW database.

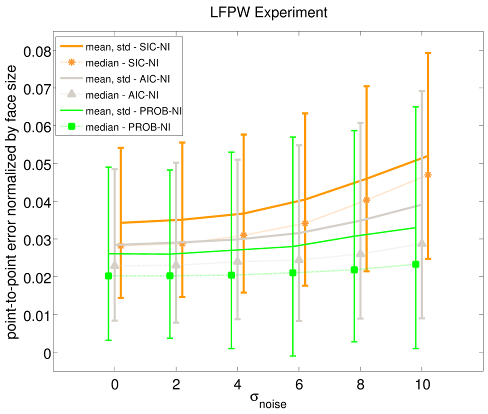

Figure 4: Fitting statistics obtained by initializing with different values of σ_{noise} on the LFPW database.

5.2. Comparison with state-of-the-art methods

In this experiment we tested the performance of our Bayesian AAM formulation, against the Robust Discriminative Response Map Fitting (DRMF) for CLMs of [9] and the Supervised Descent Method (SDM) of [11].

As before, all AAM algorithms were trained on the available 813 training images of the LFPW dataset. Results for [11] and [9] were directly obtained using the code and models provided by the authors which can be downloaded from [31] and [32] respectively (note that these models were potentially trained using thousands of images, in comparison to the only 813 images used to trained our models). Results are reported on the 330 testing images of the Helen [27] database and on the entire 337 images of the AFW [8]

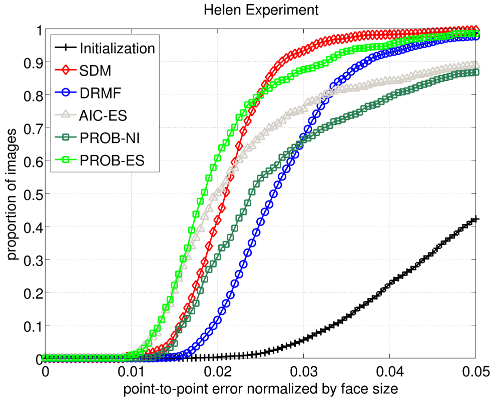

Figure 5: CED curves on the Helen database.

Method	Median	Mean	Std
SDM	0.0209	0.0216	**0.0059**
DRMF	0.0265	0.0280	0.0086
AIC-ES	0.0199	0.0274	0.0211
PROB-NI	0.0238	0.0394	0.2980
PROB-ES	**0.0184**	**0.0209**	0.0093

Table 1: Fitting statistics for the Helen databse.

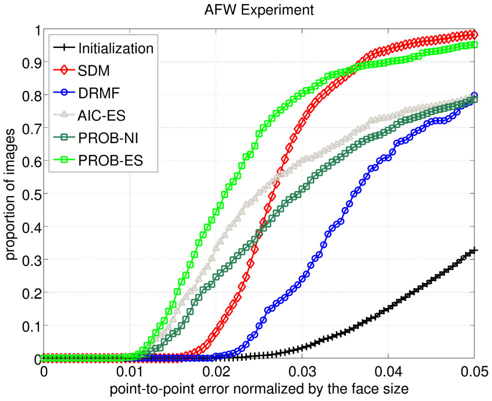

Figure 6: CED curves on the AFW database.

Method	Median	Mean	Std
SDM	0.0265	0.0273	0.0507
DRMF	0.0363	0.0517	0.0611
AIC-ES	0.0250	0.0375	0.0323
PROB-NI	0.0296	0.0796	1.1577
PROB-ES	**0.0212**	**0.0245**	**0.0132**

Table 2: Fitting statistics for the AFW database.

database. Ground truth annotations for both databases were again downloaded from [30]. Note that, compared to the previous LFPW, the images of Helen and specially of AFW appear to be much more natural and rich in variations and, consequently, are even more difficult to fit.

Both [9] and [11] take full advantage of powerful non-linear image features (*i.e.* HoG and SIFT, respectively) to achieve state-of-the-art results. For this reason, in this experiment, we use the Edge Structure (ES) features proposed in [33] as the texture representation used to build the texture model of our approach (the use of HoG and SIFT features in AAMs has never been investigated in the existent literature and lies out of the scope of this paper). For the sake of completeness, we also report results for our method using normalized intensities and for AIC using the same edge structure features.

Results for these experiments are shown in Fig. 6 and Fig. 5 and Table 1 and Table 2 (for visual inspection please see the fitting results in Fig. 7 and our supplementary material). We report CED curves and fitting statistics obtained by initializing all methods using the bounding box initializations provided by [30], which were obtained using the face detector of [8]. The results show that our approach (PROB-ES) achieves state-of-the-art results in both databases, largely outperforming the DRMF and performing marginally better than SDM (our approach is more accurate but slightly less robust). We find this results remarkable, specially considering that our Bayesian approach was trained using only 811 images in comparison to the potentially thousands of images used to trained the DRMF and SDM methods.

6. Conclusions

In this paper we present a novel Bayesian formulation of AAMs. In particular, by marginalizing out the latent texture space we derive a novel cost function that depends only on the shape parameters and propose a novel fitting algorithm to optimize it. We show that our Bayesian AAM formulation outperforms the most recently proposed state-of-the-art methods for facial alignment "in the wild" in two extremely challenging cross-database experiments.

Acknowledgements The work of Joan Alabort-i-Medina was funded by a DTA studentship from Imperial College London and by the Qualcomm Innovation Fellowship. The work of Stefanos Zafeiriou was partially funded by the EPSRC project EP/J017787/1 (4DFAB).

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001. 1, 2, 3

[2] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, 2004. 1, 2, 3, 5, 6

[3] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH*, 1999. 1

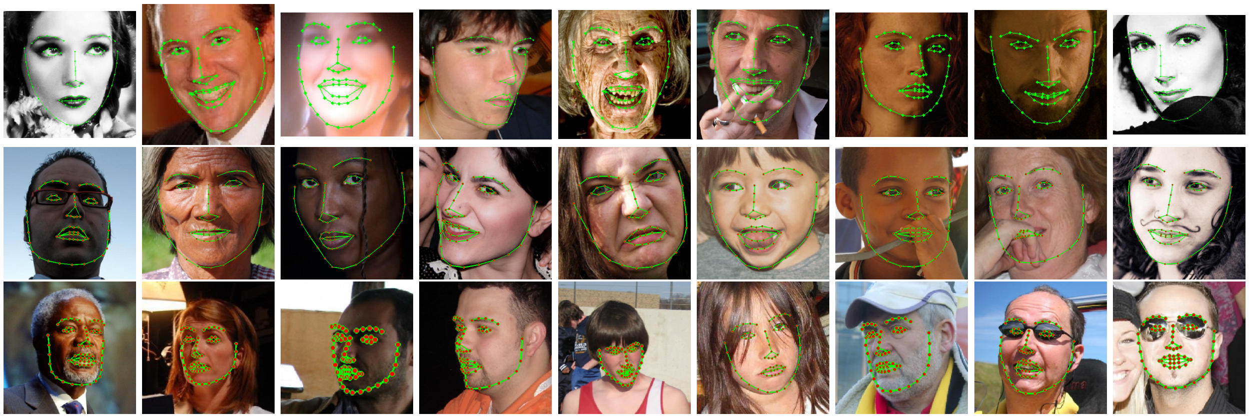

Figure 7: Selected results from our method on the 3 databases: LFPW (first row) (using normalized pixel intensities as image representation), Helen (second row) and AWF (third row) (both using edge structure features as image representation).

[4] G. Papandreou and P. Maragos, "Adaptive and constrained algorithms for inverse compositional active appearance model fitting," in *CVPR*, 2008. 1, 2, 3, 4, 5, 6

[5] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image and Vision Computing*, 2005. 2, 3, 4, 5, 6

[6] B. Amberg, A. Blake, and T. Vetter, "On compositional image alignment, with an application to active appearance models," in *CVPR*, 2009. 2, 3

[7] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, 2011. 2

[8] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012. 2, 5, 6, 7

[9] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *CVPR*, 2013. 2, 5, 6, 7

[10] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *CVPR*, 2012. 2

[11] Xuehan-Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *CVPR*, 2013. 2, 5, 6, 7

[12] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1999. 2, 3

[13] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1997. 2, 3, 4

[14] S. Roweis, "Em algorithms for pca and spca," *Advances in neural information processing systems*, 1998. 2, 3

[15] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, 1991. 2

[16] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. MIT Press, 2006. 3

[17] Y. Guan and J. G. Dy, "Sparse probabilistic principal component analysis," in *AISTATS*, 2009. 3

[18] Z. Khan and F. Dellaert, "Robust generative subspace modeling: The subspace t distribution," 2004. 3

[19] A. U. Batur and M. H. Hayes, "Adaptive active appearance models," *IEEE Transactions on Image Processing*, 2005. 3

[20] X. Liu, "Discriminative face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. 3

[21] J. Saragih and R. Göcke, "Learning aam fitting through simulation," *Pattern Recognition*, 2009. 3

[22] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, 2004. 3

[23] S. Baker, R. Gross, and I. Matthews, "Lucas-kanade 20 years on: A unifying framework: Part 3," CMU - Robotics Institute, Tech. Rep. CMU-RI-TR-03-35, 2003. 5

[24] L. van der Maaten and E. Hendriks, "Capturing appearance variation in active appearance models," in *CVPR-W*, 2010. 5

[25] T. Cootes, K. Walker, and C. Taylor, "View-based active appearance models," in *FG*, 2000. 5

[26] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *CVPR*, 2011. 5, 6

[27] V. Le, J. Brandt, Z. Lin, L. D. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *ECCV*, 2012. 5, 6

[28] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *CVPR-W*, 2013. 6

[29] ——, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *ICCV-W*, 2013. 6

[30] http://ibug.doc.ic.ac.uk/resources/300-W/. 6, 7

[31] http://www.humansensing.cs.cmu.edu/intraface/. 6

[32] https://sites.google.com/site/akshayasthana/clm-wild-code. 6

[33] T. Cootes and C. Taylor, "On representing edge structure for model matching," in *CVPR*, 2001. 7