Deformable Face Fitting with Soft Correspondence Constraints

Jason M. Saragih, Simon Lucey, Jeffrey F. Cohn
The Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213, USA
{jsaragih,slucey,jeffcohn}@cs.cmu.edu

Abstract

Despite significant progress in deformable model fitting over the last decade, the problem of efficient and accurate person-independent face fitting remains a challenging problem. In this work, a reformulation of the generative fitting objective is presented, where only soft correspondences between the model and the image are enforced. This has the dual effect of improving robustness to unseen faces as well as affording fitting time which scales linearly with the model's complexity. This approach is compared with three state-of-the-art fitting methods on the problem of person independent face fitting, where it is shown to closely approach the accuracy of the currently best performing method while affording significant computational savings.

1. Introduction

Deformable objects such as the human face are often parameterized using separate linear models of shape and appearance. Prototypes utilizing this parameterization include the active shape model (ASM) [7], the active appearance model (AAM) [6] and the 3D morphable model (3DMM) [5]. The various prototypes are designed to handle specific kinds of visual objects. The ASM is best suited for objects with strong edge features and the AAM for objects that require a dense appearance representation. The 3DMM extends the AAM's application domain to 2.5D visual objects (i.e. 2D surface embedded in 3D). For each of these prototypes, the aim of fitting is to find the model parameters that best describe the visual object in an image. In machine learning this type of approach is often referred to as generative as it adopts an *analysis-by-synthesis* strategy.

Generative methods present a principled approach to deformable face fitting. However, most current formulations suffer from two main drawbacks. Firstly, efficient parameter updates cannot be attained without approximations due to the coupling of shape and appearance parameters within the formulated objective. Secondly, generalization is limited due to the inability of the model to synthesize the whole gamut of appearance variations exhibited by complex visual objects. As discussed in [2], ignoring peculiarities about the specific problem of deformable face fitting and treating it as a generic function minimization problem often leads to the inefficient Lucas-Kanade (LK) method [11]. Earlier methods, such as [6], use a fixed approximation for LK's Jacobian that results in a rapid fitting procedure. More recently, a number of methods [1, 12, 16] have been proposed that reformulate the optimization procedure using the inverse-compositional paradigm [10]. By reversing the roles of the image and appearance model, components of the updates pertaining to the warp can be precomputed. However, without making further approximations, the optimization procedure can still be computationally expensive [1].

Typical generative methods also suffer from limited generalizability [8, 14]. As a result of its linear parameterization, the model lacks the capacity to compactly represent the whole gamut of appearance variations of a complex object, such as in the person-independent fitting problem. As such, the objective deployed in generative fitting, typically formulated as least squares, is often unduly effected by the unmodeled appearance.

In this work we address the two main drawbacks of typical generative fitting described above, presenting a formulation that applies to both 2D and 2.5D models. The main insight is that when only *soft* correspondences between the model and the image are enforced, then observation uncertainty is shared between the shape and appearance. This has the effect that unseen appearance variabilities are better handled, leading to improved generalization. Rapid fitting is achieved through a mixed inverse-compositional-forward-additive parameter update scheme, resulting in a computational complexity that scales linearly with the model's complexity. In Section 2, an overview of the parameterization and fitting of deformable face models is presented. Our fitting formulation is presented in Section 3 and quantitatively evaluated in Section 4. Section 5 concludes with an overview and mention of future work.

2. Generative Face Models

A review of generative face models is presented in this section. Their parameterization is described in Section 2.1, where the nomenclature adopted in this paper is outlined. An overview of generative fitting and their current limitations is presented in Section 2.2.

2.1. Parameterization

Intrinsic (local) sources of shape and appearance variations are commonly parameterized linearly as[1]:

$$\mathcal{S}_l(\mathbf{p}_s)\colon \Re^{M_s} \rightarrow \Re^{Dn} = \boldsymbol{\mu}_s + \boldsymbol{\Phi}_s \mathbf{p}_s \tag{1}$$

$$\mathcal{A}_l(\mathbf{p}_a)\colon \Re^{M_a} \rightarrow \Re^{PN} = \boldsymbol{\mu}_a + \boldsymbol{\Phi}_a \mathbf{p}_a, \tag{2}$$

where n, D, N, P, M_s and M_a respectively denote the number of points defining shape, shape dimensionality (2D or 3D), number of pixels in appearance, number of image planes, number of shape modes and the number of appearance modes. Here, $\{\boldsymbol{\mu}_s, \boldsymbol{\mu}_a\}$ and $\{\boldsymbol{\Phi}_s, \boldsymbol{\Phi}_a\}$ denote the mean and basis (modes) of variation respectively.

The appearance of a visual object in the image is often effected by extrinsic (global) sources of variation, both geometrically and photometrically. It is common practice to model extrinsic photometric variations also as a linear model [1, 6]:

$$\mathcal{A}_g(\mathbf{a}; \mathbf{g}_a)\colon \Re^P \times \Re^{2P} \rightarrow \Re^P = \boldsymbol{\alpha} \odot \mathbf{a} + \boldsymbol{\beta}, \tag{3}$$

where $\mathbf{g}_a = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ are the gain and bias for each image channel. Extrinsic geometric variations are more involved as they typically represent the object's rigid motion:

$$\mathcal{S}_g(\mathbf{s}; \mathbf{g}_s)\colon \Re^{Dn} \times \Re^{G} \rightarrow \Re^{Dn} = s(\mathbf{I} \otimes \mathbf{R})\mathbf{s} + \mathbf{1} \otimes \mathbf{t}, \tag{4}$$

where $\mathbf{g}_s = \{s, \mathbf{R}, \mathbf{t}\}$, $\mathbf{R}$ is a $(D \times D)$ rotation matrix, $\mathbf{t}$ is a translation, s is a scaling factor and G denotes the number of parameters defining the rigid motion.

Together, the local and global transformations constitute the visual object's generative model:

$$\mathcal{S}(\mathbf{p}_s, \mathbf{g}_s) = \mathcal{S}_g(\Diamond; \mathbf{g}_s) \circ \mathcal{S}_l(\mathbf{p}_s) \tag{5}$$

$$\mathcal{A}(\mathbf{x}; \mathbf{p}_a, \mathbf{g}_a) = \mathcal{A}_g(\Diamond; \mathbf{g}_a) \circ \mathcal{A}_l(\mathbf{x}; \mathbf{p}_a), \tag{6}$$

Where we have used $\mathcal{A}(\mathbf{x}; \mathbf{p}_a)$ to denote the appearance at pixel location $\mathbf{x}$ in the model frame.

[1] **Notation:** Vectors are written in lowercase bold and matrices in uppercase bold, where $\mathbf{1}$ denotes the all one vector and $\mathbf{I}$ the identity matrix. Greek letters denote either vectors or matrices depending on context. The Hadamard (element wise) and Kronecker (tiling) products are written as $\odot$ and $\otimes$, respectively. The diag$\{\mathbf{x}\}$ operator makes a diagonal matrix with the components of $\mathbf{x}$ as its diagonal entries. Functions are written in upper case calligraphic font with $\circ$ denoting their composition. When composing functions with multiple parameters, $\Diamond$ denotes the parameters resulting from the output of the composed function, for example: $\mathcal{A}(\mathcal{B}(x); y) = \mathcal{A}(\Diamond; y) \circ \mathcal{B}(x)$.

2.2. Fitting

The objective of generative fitting is to minimize a cost function of the form:

$$\mathcal{E}(\mathcal{I}; \boldsymbol{\theta}) = \underbrace{\mathcal{D}(\mathcal{I}; \boldsymbol{\theta})}_{\text{Data term}} + \underbrace{\lambda_s \mathcal{R}(\mathbf{p}_s) + \lambda_a \mathcal{R}(\mathbf{p}_a)}_{\text{Regularization term}}, \tag{7}$$

where $\mathcal{I}$ is the image and $\boldsymbol{\theta} = \{\mathbf{p}_s, \mathbf{g}_s, \mathbf{p}_a, \mathbf{g}_a\}$. The data term $\mathcal{D}$ is often set as the least squares error between the cropped image and the model's appearance:

$$\sum_{i=1}^{N} \left\| \mathcal{I} \circ \mathcal{W}(\mathbf{x}_i; \Diamond) \circ \mathcal{P} \circ \mathcal{S}(\mathbf{p}_s, \mathbf{g}_s) - \mathcal{A}(\mathbf{x}_i; \mathbf{p}_a, \mathbf{g}_a) \right\|^2, \tag{8}$$

where $\{\mathbf{x}_i\}_{i=1}^N$ is the set of locations in the model frame defining appearance, $\mathcal{P}$ is a projection onto the image and $\mathcal{W}$ is a warping function, often chosen as the piecewise affine warp [6, 12, 16]:

$$\mathcal{W}(\mathbf{x}_i; \mathbf{s})\colon \Re^{Dn} \rightarrow \Re^D = \mathbf{W}_i^{(D \times Dn)} \mathbf{s}\,. \tag{9}$$

Note that we parameterize the warp using its target *nodes*, similar to that derived in [3] for the thinplate spline.

The regularization $\mathcal{R}_s$ and $\mathcal{R}_a$, which correspond to priors over the shape and appearance parameters, are often chosen as anisotropic Gaussians [4, 15]:

$$\mathcal{R}_s(\mathbf{p}_s) = \mathbf{p}_s^T \text{diag}\left\{\boldsymbol{\sigma}_s^{-2}\right\} \mathbf{p}_s, \tag{10}$$

where $\boldsymbol{\sigma}_s^2$ denotes the variances along the directions of shape variation (similarly for appearance). However, many works, for example [6, 7, 12], do away with a regularization term, solving the maximum-likelihood (ML) instead of the maximum-*a-posteriori* (MAP) problem.

Utilizing a generic optimization strategy such as Gauss-Newton to minimize the data term in Equation (8) leads to expensive updates for $\boldsymbol{\theta}$, since the derivative of the cropped image with respect to the shape parameters needs to be recomputed at each iteration [1]. As such, in inverse-compositional based methods, at each iteration the solution is sought for the reformulated problem:

$$\sum_{i=1}^{N} \left\| \mathcal{I} \circ \mathcal{W}(\mathbf{x}_i; \mathbf{s}) - \mathcal{A}(\Diamond; \mathbf{p}_a, \mathbf{g}_a) \circ \mathcal{W}(\mathbf{x}_i; \Delta\mathbf{s}) \right\|^2, \tag{11}$$

optimizing over $\{\Delta\mathbf{p}_s, \Delta\mathbf{g}_s, \mathbf{p}_a, \mathbf{g}_t\}$, where:

$$\mathbf{s} = \mathcal{P} \circ \mathcal{S}(\mathbf{p}_s, \mathbf{g}_s) \quad \text{and} \quad \Delta\mathbf{s} = \mathcal{P} \circ \mathcal{S}(\Delta\mathbf{p}_s, \Delta\mathbf{g}_s). \tag{12}$$

The forward warp is then found by inverting the update warp and composing it with the current estimate.

Since $\boldsymbol{\mu}_a$ and $\boldsymbol{\Phi}_a$ are known and the derivative of the warp is always evaluated at the identity, the Jacobian of $\boldsymbol{\mu}_a$ and $\boldsymbol{\Phi}_a$

can be precomputed. However, the Jacobian of the synthesized appearance is not fixed as it depends on $\{\mathbf{p}_a, \mathbf{g}_a\}$ [1]. As such, when utilizing a Gauss-Newton step to optimize Equation (11), the Jacobian and its pseudo-inverse (linear update model) must be rebuilt, the computational cost of which scales cubically with the model's complexity. In [16], fixed linear updates were attained by applying $\mathcal{W}(\mathbf{x}; \Delta\mathbf{s})$ only to $\boldsymbol{\mu}_a$, rather than to $\boldsymbol{\Phi}_a$ as well. However, this is equivalent to approximating the appearance Jacobian with that evaluated at the identity appearance, an instance of the efficient approximation to the simultaneous inverse compositional method [1].

The project-out method [12] affords extremely rapid fitting by minimizing in the subspace orthogonal to $\boldsymbol{\Phi}_a$:

$$\|\mathcal{I} \circ \mathcal{W}(\mathbf{s}) - \boldsymbol{\mu}_a \circ \mathcal{W}(\Delta\mathbf{s})\|^2_{\text{span}(\boldsymbol{\Phi}_a)^\perp} \ , \qquad (13)$$

resulting in fixed linear updates. However, as shown in [8], the efficacy of this approach greatly deteriorates as the model's complexity increases. Although some possible causes of this deterioration was given in [8], a satisfactory explanation is still lacking. Examining the form presented in Equation (13), however, the reason for the performance deterioration becomes apparent: the model frame warp $\mathcal{W}(\Delta\mathbf{s})$ is applied only to $\boldsymbol{\mu}_a$ rather than to $\boldsymbol{\Phi}_a$ as well. A more appropriate objective is to minimize the template-to-image difference in $\text{span}(\boldsymbol{\Phi}_a \circ \mathcal{W}(\Delta\mathbf{s}))^\perp$. However, this ruins the precomputability of the updates. Therefore, the project-out method makes the same approximation as the efficient approximation to the simultaneous inverse-compositional method [1].

Apart from difficulty in attaining efficient fitting without approximations, another weakness of current generative fitting methods is their limited generalizability. In [8, 14], investigations into this aspect of face fitting found that the main source of difficulty lies in the inability of the appearance model to accurately generate previously unseen appearance. This problem is exacerbated when the model exhibits large shape variability, as it tends to deform, often away from the true shape, to account for the unmodeled appearance. This results because the unmodeled appearance does not generally follow an isotropic Gaussian distribution, which is implicitly assumed in the least squares fitting objective. It was found in [14] that optimal performance was attained when full appearance and 60% shape variance is retained. However, this can have the effect that the shape model lacks expressiveness as well as a significant increase in fitting time due to the highly complex appearance model used.

3. A Soft Correspondence Formulation

To address the main difficulties associated with generative fitting, we propose reformulating the fitting objective such that only *soft* correspondences between the model's shape and image coordinates are enforced. For this, a set of auxiliary variables, $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N$, are introduced that denote locations in the image frame corresponding to each pixel in the model. Inspired by Thirion's demons algorithm [18] for optical flow, the idea is to let undesired shape deformations, stemming from unmodeled appearance, to be handled by the correspondences, without excessively disturbing the model's shape. As an added benefit, with the formulation presented shortly, these correspondences decouple components of the fitting objective pertaining to shape and appearance, allowing efficient *fixed* updates for the intrinsic parameters to be afforded without approximations.

Let us redefine the data term in Equation (8) as:

$$\mathcal{D}(\mathcal{I}; \boldsymbol{\theta}) = \mathcal{F}(\mathcal{I}; \mathbf{Z}, \mathbf{p}_a, \mathbf{g}_a) + \lambda_c \, \mathcal{C}(\mathbf{Z}; \mathbf{p}_s, \mathbf{g}_s). \qquad (14)$$

Here, $\mathcal{F}$ denotes a measure of fit between the model's appearance and the image, defined at $\mathbf{Z}$:

$$\mathcal{F} = \sum_{i=1}^{N} \|\mathcal{A}_g(\Diamond; \mathbf{g}_a) \circ \mathcal{I} \circ \mathcal{P}(\mathbf{z}_i) - \mathcal{A}_l(\mathbf{x}_i; \mathbf{p}_a)\|^2. \qquad (15)$$

The term $\mathcal{C}$ penalizes deviations of the correspondences from locations defined by the model's shape:

$$\mathcal{C} = \sum_{i=1}^{N} \|\mathcal{S}_g(\mathbf{z}_i; \mathbf{g}_s) - \mathcal{W}(\mathbf{x}_i; \Diamond) \circ \mathcal{S}_l(\mathbf{p}_s)\|^2 . \qquad (16)$$

In this work, it is assumed that $\mathcal{W}$ is a linear function of the destination shape, which define the *nodes* of the warp as in Equation (9). Warps that exhibit this form include the piecewise affine and thinplate spline warps. In our formulation, the warp is used in a slightly different way from the convention of warping the model to image coordinates for the purposes of appearance cropping. Here, the warp is used to transform the model's shape to another shape of the same dimensionality. In the 2D case, no distinction is made between the typical use of warp and how it is used here. However, in the 2.5D case, the warp transforms 3D locations to other 3D locations, which can also be defined linearly. In the exposition that follows, we define:

$$\mathcal{W} \circ \mathcal{S}_l(\mathbf{p}_s) = \mathbf{W} \, (\boldsymbol{\mu}_s + \boldsymbol{\Phi}_s \, \mathbf{p}_s) = \hat{\boldsymbol{\mu}}_s + \hat{\boldsymbol{\Phi}} \, \mathbf{p}_s \ , \qquad (17)$$

where rows of $\mathbf{W}$ are set to $\{\mathbf{W}_i\}_{i=1}^N$. When a dense shape model is used, as in 3DMM's, $\mathbf{W}$ is the identity.

Examining Equations (15) and (16) one immediately notices that the extrinsic geometric and photometric transformations are applied to the image rather than the model frame components. Optimizing such an objective involves a mixture of forward-additive and inverse-compositional updates for the intrinsic and extrinsic parameters respectively.

In order to allow such use, the extrinsic transformations must form a group. The transformations defined in Equations (3) and (4) exhibit this property. Although this choice may appear unconventional, it will be shown in Section 3.2 that such a formulation has significant impact on the computational efficiency of the method.

Intuitively, the choice of the terms in Equations (15) and (16) have the effect of favoring $\mathbf{Z}$ that simultaneously defines locations in $\mathcal{I}$ that best fit the generated model's appearance while also adhering to the space of allowable geometric deformations. As λ_c in Equation (14) increases, $\mathbf{Z}$ becomes increasingly constrained to adhere to the shape model. As $\lambda_c \to \infty$, the correspondences become deterministic, yielding the original LK algorithm (albeit with inverted extrinsic transformations). As such, this formulation allows a level of "slackness" to the shape and appearance fit, with the trade-off between them regulated by λ_c. This slackness has the effect of sharing observation uncertainties between the shape and appearance.

With the reformulation of the data term in Equation (14), optimization is now required over an extra DN parameters compared to typical generative fitting scenarios. As such, a simultaneous optimization strategy is computationally expensive. For this, we use a parallel axis optimization strategy, where parameters are iteratively grouped and optimized separately. Here, we propose partitioning the parameters into the correspondences $\mathbf{Z}$, the local parameters $\{\mathbf{p}_s, \mathbf{p}_a\}$ and the global parameters $\{\mathbf{g}_s, \mathbf{g}_a\}$.

3.1. Optimizing the Correspondences

Keeping the other parameters fixed, the correspondences are found by minimizing:

$$\mathcal{E}(\mathbf{Z}) = \sum_{i=1}^{N} \left\| \boldsymbol{\alpha} \odot \mathcal{I} \circ \mathcal{P}\left(\mathbf{z}_i\right) + \boldsymbol{\beta} - \mathcal{A}_l(\mathbf{x}_i; \mathbf{p}_a) \right\|^2 + \lambda_c \sum_{i=1}^{N} \left\| s\, \mathbf{R}\, \mathbf{z}_i + \mathbf{t} - \mathcal{W}(\mathbf{x}_i; \lozenge) \circ \mathcal{S}_l(\mathbf{p}_s) \right\|^2 . \quad (18)$$

Since the spatial locations within an image are generally related to its pixel values nonlinearly, this problem constitutes a nonlinear function over $\mathbf{Z}$. However, if we assume that the current estimates of the correspondences are close to their optimal settings, then a first order Taylor expansion of the image is a reasonable assumption:

$$\mathcal{I} \circ \mathcal{P}\left(\mathbf{z}_i\right) \approx \underbrace{\mathcal{I} \circ \mathcal{P}\left(\mathbf{z}_i^c\right)}_{\mathcal{I}_i} + \underbrace{\nabla \mathcal{I} \frac{\partial \mathcal{P}}{\partial \mathbf{z}_i}}_{\nabla \mathcal{I}_i} \left[\mathbf{z}_i - \mathbf{z}_i^c\right], \quad (19)$$

where $\mathbf{z}_i^c$ is the current estimate of $\mathbf{z}_i$. Note that when the shape model is 2D, $\mathcal{P}$ is simply the identity function, and when a 3D shape model is used with a weak perspective projection, $\mathcal{P}$ simply extracts the x and y components of $\mathbf{z}$. A full perspective 3D model is more involved and is out of the scope of this paper, however, the formulation here allows a direct derivation to be made, modifying only $\frac{\partial \mathcal{P}}{\partial \mathbf{z}_i}$.

With this linearization, the cost function in Equation (18) becomes quadratic. Furthermore, since the correspondences exits in separate terms of the summation, the cost function for each takes the form:

$$\mathcal{E}(\mathbf{z}_i) = \left\| \boldsymbol{\alpha} \odot \left(\nabla \mathcal{I}_i\, \mathbf{z}_i\right) + \delta \mathbf{a}_i \right\|^2 + \lambda_c \left\| s \mathbf{R} \mathbf{z}_i + \delta \mathbf{s}_i \right\|^2, \quad (20)$$

where:

$$\delta \mathbf{s}_i = \mathbf{t} - \mathcal{W}(\mathbf{x}_i; \lozenge) \circ \mathcal{S}_l(\mathbf{p}_s) \quad (21)$$

$$\delta \mathbf{a}_i = \boldsymbol{\alpha} \odot \left(\mathcal{I}_i - \nabla \mathcal{I}_i\, \mathbf{z}_i^c\right) + \boldsymbol{\beta} - \mathcal{A}_l(\mathbf{x}_i; \mathbf{p}_a) . \quad (22)$$

The solution, then, is given by:

$$\mathbf{z}_i = -\mathbf{H}_i^{-1} \left[\nabla \mathcal{I}_i^T (\boldsymbol{\alpha} \odot \delta \mathbf{a}_i) + \lambda_c\, s\, \mathbf{R}^T \delta \mathbf{s}_i\right], \quad (23)$$

where:

$$\mathbf{H}_i = \nabla \mathcal{I}_i^T \mathrm{diag}\{\boldsymbol{\alpha}^2\} \nabla \mathcal{I}_i + \lambda_c\, s^2\, \mathbf{R}^T \mathbf{R} . \quad (24)$$

This is a $(D \times D)$ linear system that affords an efficient evaluation for each $\mathbf{z}_i$. It constitutes the constrained optical flow estimate at each pixel, where the color constancy equation is defined between the image and the locally synthesized appearance. With no geometric constraint, $\mathbf{H}_i$ is rank deficient. However, the shape constraint in the model frame ensure that $\mathbf{H}_i$ is invertible (i.e. $\{\lambda_c, s\} > 0$ and $\mathbf{R}^T \mathbf{R}$ is positive definite).

3.2. Optimizing the Intrinsic Parameters

By virtue of the additional variables $\mathbf{Z}$, the components of Equation (14) pertaining to shape and appearance are decoupled from each other. As such, simultaneously optimizing over the local shape and appearance parameters is equivalent to optimizing each independently. Furthermore, for fixed $\{\mathbf{Z}, \mathbf{g}_s, \mathbf{g}_a\}$, the objective is quadratic.

With the linearization in Equation (19) and letting:

$$\nabla \mathbf{s} = s\, (\mathbf{I} \otimes \mathbf{R})\, [\mathbf{z}_1; \ldots; \mathbf{z}_N] + \mathbf{1} \otimes \mathbf{t} - \hat{\boldsymbol{\mu}}_s \quad (25)$$

$$\nabla \mathbf{a} = \begin{bmatrix} \boldsymbol{\alpha} \odot (\mathcal{I}_1 + \nabla \mathcal{I}_1 [\mathbf{z}_1 - \mathbf{z}_1^c]) \\ \vdots \\ \boldsymbol{\alpha} \odot (\mathcal{I}_N + \nabla \mathcal{I}_N [\mathbf{z}_N - \mathbf{z}_N^c]) \end{bmatrix} + \mathbf{1} \otimes \boldsymbol{\beta} - \boldsymbol{\mu}_a , \quad (26)$$

the cost function pertaining to intrinsic shape and appearance parameters are respectively given by:

$$\mathcal{E}(\mathbf{p}_s) = \left\| \nabla \mathbf{s} - \hat{\boldsymbol{\Phi}}_s \mathbf{p}_s \right\|^2 + \frac{\lambda_s}{\lambda_c} \mathbf{p}_s^T \mathrm{diag}\left\{\boldsymbol{\sigma}_s^{-2}\right\} \mathbf{p}_s \quad (27)$$

$$\mathcal{E}(\mathbf{p}_a) = \left\| \nabla \mathbf{a} - \boldsymbol{\Phi}_a \mathbf{p}_a \right\|^2 + \lambda_a \mathbf{p}_a^T \mathrm{diag}\left\{\boldsymbol{\sigma}_a^{-2}\right\} \mathbf{p}_a , \quad (28)$$

where we have assumed Gaussian priors on the parameters. Solutions for the intrinsic parameters are given by:

$$\mathbf{p}_s = \left(\hat{\mathbf{\Phi}}_s^T \hat{\mathbf{\Phi}}_s + \frac{\lambda_s}{\lambda_c} \text{diag}\left\{ \boldsymbol{\sigma}_s^{-2} \right\} \right)^{-1} \hat{\mathbf{\Phi}}_s^T \, \nabla \mathbf{s} \qquad (29)$$

$$\mathbf{p}_a = \left(\mathbf{\Phi}_a^T \mathbf{\Phi}_a + \lambda_a \, \text{diag}\left\{ \boldsymbol{\sigma}_a^{-2} \right\} \right)^{-1} \mathbf{\Phi}_a^T \, \nabla \mathbf{a}. \qquad (30)$$

In these equations, all components apart from $\nabla \mathbf{s}$ and $\nabla \mathbf{a}$ can be precomputed, allowing fixed linear updates to be attained. Since the expressive power of linear models rely on the number of these local parameters, the savings here is significant, especially for more complex models, since the computational complexity increases only linearly with the number of modes of shape and appearance variation.

These fixed updates are made possible through the specific formulation of the problem used here. Firstly, the introduction of $\mathbf{Z}$ decouples the shape parameters, both from the image and the intrinsic appearance parameters. In the conventional forward additive formulation, $\mathbf{p}_s$ is coupled with the image, whereas in the inverse compositional formulation it is coupled with $\mathbf{p}_a$. It should be noted, however, that when a sparse point set is used, as in AAMs, the linear system for the shape parameters is much larger than that of typical formulations. Nonetheless, as M_s increases, the formulation here may still affords significant computational savings. For 3DMMs, where a dense point set is used, savings are attained, regardless of the model's complexity.

Secondly, applying the extrinsic photometric transformation to $\mathcal{I}$ and the extrinsic geometric transformation to $\mathbf{Z}$ decouples the intrinsic and extrinsic parameters. Without this measure, Equations (29) and (30) would depend on the current estimates of the extrinsic parameters, preventing the updates from being precomputed. It should be noted however, that in a ML framework, the photometric gain and bias can be appended to $\mathbf{\Phi}_a$, allowing fixed updates to be attained simultaneously for both local and global transformations. Similarly, for the global geometric transformation, some authors (for example, see [12]) implement a 2D similarity transform linearly by prepending four orthogonal columns to $\mathbf{\Phi}_s$. However, such a simplification cannot be made for 3D rigid motion. Furthermore, since the local and global parameters become inseparable, a Gaussian prior cannot be placed on the intrinsic parameters alone.

3.3. Optimizing the Extrinsic Parameters

As with the local parameters, the global parameters are also decoupled within the cost function. Furthermore, for multi-plane images, the photometric gain and bias for each plane are also decoupled from each other. Letting:

$$\mathbf{\Psi}_j = \begin{bmatrix} \mathcal{I}_{1(j)} + \nabla \mathcal{I}_{1(j)} [\mathbf{z}_1 - \mathbf{z}_1^c] & 1 \\ \vdots & \vdots \\ \mathcal{I}_{N(j)} + \nabla \mathcal{I}_{N(j)} [\mathbf{z}_N - \mathbf{z}_N^c] & 1 \end{bmatrix}, \qquad (31)$$

where $\mathcal{I}_{i(j)}$ denotes the j^{th} plane of the i^{th} cropped pixel, the cost function pertaining to $\{\boldsymbol{\alpha}_{(j)}, \boldsymbol{\beta}_{(j)}\}$ is:

$$\mathcal{E}(\boldsymbol{\alpha}_{(j)}, \boldsymbol{\beta}_{(j)}) = \left\| \mathbf{\Psi}_j \, [\, \boldsymbol{\alpha}_{(j)} \, ; \, \boldsymbol{\beta}_{(j)} \,] - \mathcal{A}_l(\mathbf{p}_a)_j \right\|^2, \quad (32)$$

where $\mathcal{A}_l(\mathbf{p}_a)_j$ denotes the j^{th} plane of the generated appearance. This quadratic function has the solution:

$$[\, \boldsymbol{\alpha}_{(j)} \, ; \, \boldsymbol{\beta}_{(j)} \,] = \left(\mathbf{\Psi}_j^T \mathbf{\Psi}_j \right)^{-1} \mathbf{\Psi}_j^T \mathcal{A}_l(\mathbf{p}_a)_j, \qquad (33)$$

that allows a rapid evaluation at each iteration.

The component of the error function pertaining to the extrinsic geometric transformation is given by:

$$\mathcal{E}(s, \mathbf{R}, \mathbf{t}) = \left\| s(\mathbf{I} \otimes \mathbf{R})\mathbf{Z} + \mathbf{1} \otimes \mathbf{t} - \mathcal{W} \circ \mathcal{S}_l(\mathbf{p}_s) \right\|^2. \qquad (34)$$

This is the extended Procrustes alignment problem [17] that also affords an efficient globally optimal solution.

3.4. Discussion

An outline of the proposed fitting algorithm is presented in Algorithm 1. At first glance, the iterative loop (steps 3 through 9) seem to resemble typical feature based fitting [4, 7]. However, those methods re-estimate the correspondences from the shape defined locations at each iteration. As such, they fail to minimize a consistent global objective between iterations, often leading to non-convergent cyclic behavior. In contrast, the method proposed here treats the correspondences as additional variables and minimizes a consistent global objective.

Finally, since the updates for $\mathbf{Z}$ constitute a constrained optical flow estimate, their predictive region is limited. To capture large deformations, the procedure should be implemented on a Gaussian pyramid, where higher pyramid levels predict large deformations, with increasingly localized predictions as the procedure descends the pyramid.

4. Experimental Evaluation

Experiments were performed on a subset of the CMU Pose, Illumination, and Expression Database (MultiPIE) [9], pertaining to ambient lighting with frontal and pseudo-frontal views. The selected subset was further partitioned into training and test sets, consisting of 100 and 239 subjects respectively, for a total of 3051 images. Also, a separate set

Algorithm 1 Fitting with Soft Correspondence Constraints

Require: $\mathcal{I}$, $\{\mathbf{p}_s, \mathbf{g}_s, \mathbf{p}_a, \mathbf{g}_a\}$, and N_i.

Initialize correspondences: $\mathbf{Z} = \mathcal{W} \circ \mathcal{S}(\mathbf{p}_s, \mathbf{g}_s)$.
$\mathbf{g}_s \leftarrow \mathcal{S}_g(\mathbf{g}_s)^{-1}$ and $\mathbf{g}_a \leftarrow \mathcal{A}_g(\mathbf{g}_a)^{-1}$.
for $i = 1$ to N_i **do**
  Linearize image {Eqn. (19)}.
  Update correspondences $\mathbf{Z}$ {Eqn. (23)}.
  Update $\{\mathbf{p}_s, \mathbf{p}_a\}$ {Eqn's. (29) and (30)}.
  Update $\mathbf{g}_a$ {Eqn. (33)} and $\mathbf{g}_s$ {see [17]}.
  Check for convergence.
end for
$\mathbf{g}_s \leftarrow \mathcal{S}_g(\mathbf{g}_s)^{-1}$ and $\mathbf{g}_a \leftarrow \mathcal{A}_g(\mathbf{g}_a)^{-1}$.
return $\{\mathbf{p}_s, \mathbf{g}_s, \mathbf{p}_a, \mathbf{g}_a\}$

of 10 subjects was used to select the regularization parameters $\{\lambda_c, \lambda_s, \lambda_a\}$ through cross validation. The database includes 68-point manual annotations for all images, allowing a quantitative analysis of performance to be made. On this database, two sets of experiments were performed. The first, which is presented in Section 4.1, was designed to investigate the effects of varying model complexity on the *best* fitting accuracy, similar to that in [14]. Performance in this set of experiments gives an indication of how well the method generalizes to unseen faces. Given the large number of additional variables, a serious concern regarding the efficacy of the proposed method is its susceptibility towards local minima. The second set of experiments, which is presented in Section 4.2, was designed to investigate this aspect of fitting, given realistic initial conditions.

In order to evaluate the 2.5D variant of our proposed method, 3D shape models were learned by applying non-rigid Structure-from-Motion [19] on the training set in each experiment. As with most annotations aimed at building 2D shape models, the topology around the periphery of the face is not preserved by the annotations in MultiPIE. Points corresponding to the inner jaw in the frontal view are annotated on the cheek in the half profile view, as illustrated in Figure 1. Including images with extreme poses can lead to unrealistic 3D shape models being learned. As such, they have been excluded from the experiments presented here.

4.1. Investigating Generalization

To investigate the robustness of the proposed approach to unseen sources of variability, shape and appearance models were first built using an increasing number of subjects selected from the training set. Then, starting from their optimal settings in each test image, the model was fit until convergence or a maximum of $N_i = 100$ iterations was performed. Perturbations from their optimal settings gives a measure of how well the models generalize under the

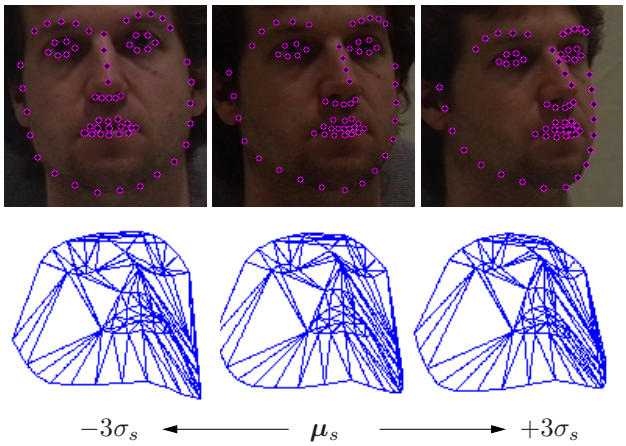

Figure 1. Effects of learning a 3D model from 2D annotations. **First row:** Example annotations of frontal, pseudo frontal and half profile views. **Second row:** one mode of variation of the 3D shape model built from a set including half profile views. The unrealistic contortions around the inner jaw result from annotations which do not correspond to the same physical location between views.

prescribed fitting objective and optimization strategy. Results are presented in Figure 2 for the 2D and 2.5D variants. The graphs (fitting curves) show the proportion of images at which various levels of maximum perturbation was exhibited, measured as the root-mean-squared (RMS) error between the annotations and the projected shape.

The results show a clear trend of performance improvement as the number of training subjects increases. This trend persists even when only a small training set of 5 and 10 subjects is used. As discussed in [8, 14], the shape model requires far fewer training instances to achieve good representation capacity compared to appearance. They also identified that performance deteriorates when the *increase* in shape representation outweighs that of appearance, which typically occurs with a small sample set. This is because the shape can deform in more ways to satisfy unmodeled appearance. However, this artifact of common generative formulations is absent in the results presented here, by virtue of the soft correspondence formulation which shares the observation uncertainty between shape and appearance.

4.2. Comparisons with other Methods

To compare the proposed approach against others under realistic initial conditions, the model's translation and scale in each test image was attained from an off-the-shelf face detector and the intrinsic shape and appearance parameters were initialized to zero (i.e. mean shape and appearance). From there, the model was fit on three levels of a Gaussian pyramid until convergence or a maximum number of iterations was performed. Its performance was compared against

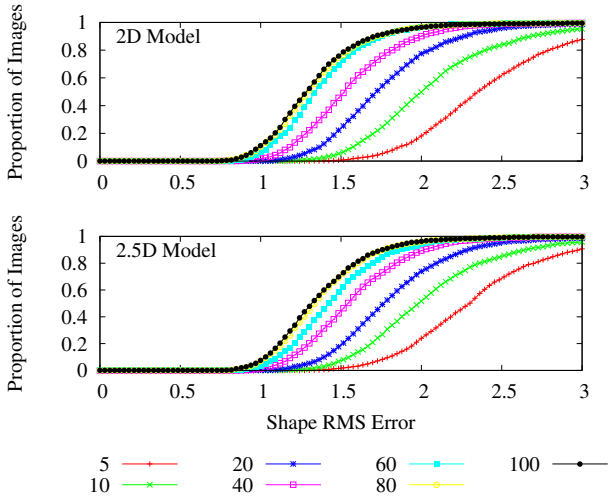

Figure 2. Effects of training set size on the optimal fitting performance on unseen faces. Legend denotes the number of subjects used to train the shape and appearance models.

three prominent generative fitting methods: the project-out inverse compositional method [12] (POIC), the simultaneous inverse compositional method [1] (SIM) and the 2D+3D method [13] (2D+3D IC). In the following, we will refer to the method proposed in this work as the Softly Constrained Correspondence method (SCC). In all cases, we set $N_i = 20$ for the POIC, SIM and 2D+3D IC methods, and set it to 100 for SCC, since it uses parallel axis optimization that requires more iterations to converge. Results of these experiments are presented in Figure 3, where the fitting time and accuracy are plotted against the total number of shape and appearance modes (i.e. $M_s + M_a$). Also shown are fitting curves, analogous to those in Figure 2, for models trained with 5 and 100 subjects. For the accuracy plot, performance is measured as the area under the fitting curve. The fitting times shown denote C++ implementations on a 1.83GHz MacBook.

From these results, one notices that although both variants of SCC outperform POIC and 2D+3D IC, they fail to outperform SIM. Furthermore, the 2.5D variant yields poorer performance than its 2D counterpart. Although SCC has the capacity to attain better accuracy as its representation capacity improves, it appears that this is not always achieved under realistic initial conditions, with the tendency to terminate in local minima. An example of this is shown in Figure 4. This is an artifact, not only of the formulation, which involves a larger set of variables, but also of the parallel axis strategy used for optimization. Since the correspondences are updated based only on local image structure and their distance from the current shape, when there is insufficient image structure their movements are highly constrained. As these correspondences are then used to update the shape

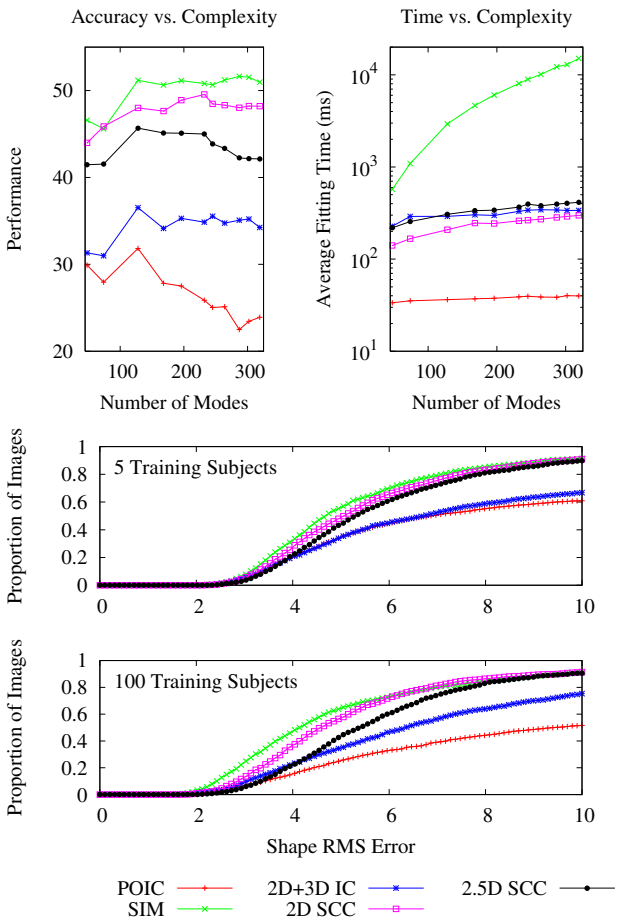

Figure 3. Performance comparison between SCC and another three methods. **Top left:** fitting accuracy, measured as the area under the fitting curve, against total number of shape and appearance modes. **Top right:** Fitting time of the various methods as model complexity varies. **Middle and bottom rows:** fitting curves of the various methods for models trained using 5 and 100 training samples respectively.

with equal weight assigned to each, undue importance is placed on correspondences with insufficient image structure. Confidence over their predictive capacity, encoded in the Hessian of Equation (24), is lost through this procedure. The problem is amplified in the 2.5D variant, which exhibits an extra N variables compared to its 2D counterpart, resulting in a more complex error terrain. Nonetheless, SCC's performance closely approaches that of SIM, and is much better than POIC or 2D+3D IC. Furthermore, it achieves a significantly reduced fitting time compared to SIM.

5. Conclusion

In this work, the typical objective for fitting generative deformable models was reformulated to enforce only soft cor-

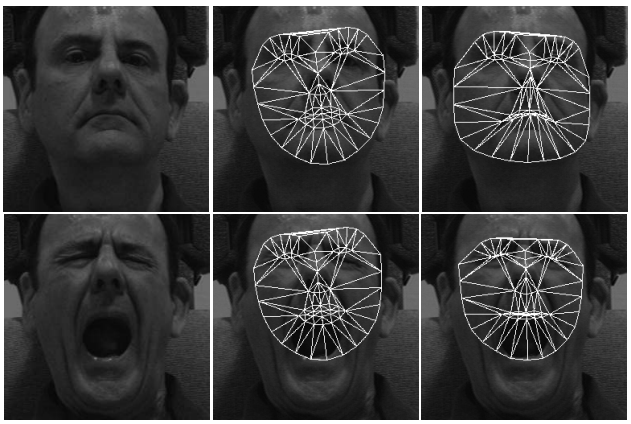

Figure 4. Some examples of fitting using the 2D SCC. **Left to right:** Area of image containing the face, initial model placement using a face detector, and final fitting result. **Top row:** Successful fitting. **Bottom row:** Fitting terminating in local minima.

respondences between the image and the model. Through experiments on the human face, it was shown that this objective has the effect of affording better fitting accuracy as the model's representation capacity improves, even when the training set is small. Furthermore, through a parallel axis optimization strategy, the computational cost of fitting is highly reduced without needing to make further approximations. However, it was also found that the method is sensitive to local minima. Despite this, performance is comparable to the currently most accurate method, while affording significant savings in computational complexity.

Future work will address the sensitivity of the method towards local minima. Reducing the number of parameters through a combined appearance representation as in [6] is a first step towards this, which also allows fixed updates for the intrinsic parameters to be attained by the formulation presented here. Performing simultaneous optimization over all parameters may also yield better robustness, where the block structure of the Hessian may be taken advantage of in order to speed up computations.

References

[1] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework: Part 3. Technical report, Robotics Institute, Carnegie Mellon University, 2003.

[2] S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework: Part 1. Technical report, Robotics Institute, Carnegie Mellon University, 2002.

[3] A. Bartoli, M. Perriollat, and S. Chambon. Generalized Thin-Plate Spline Warps. In *CVPR'07: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, June 2007.

[4] C. Basso, T. Vetter, and V. Blanz. Regularized 3D Morphable Models. In *HLK '03: Proceedings of the 1st IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis*, page 3, 2003.

[5] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. In *SIGGRAPH'99: Proceedings of the 26th International Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999.

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. In *ECCV'98: Proceedings of the 5th European Conference on Computer Vision*, volume 2, pages 484–498, 1998.

[7] T. F. Cootes and C. J. Taylor. Active Shape Models - 'Smart Snakes'. In *BMVC'92: Proceedings of the 3rd British Machine Vision Conference*, pages 266–275, 1992.

[8] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing (IVC)*, 23:1080–1093, 2005.

[9] R. Gross, I. Matthews, S. Baker, and T. Kanade. The CMU Multiple Pose, Illumination and Expression (MultiPIE) Database. Technical report, Robotics Institute, Carnegie Mellon University, 2007.

[10] G. D. Hager and P. N. Belhumeur. Efficient Region Tracking with Parametric Models of Geometry and Illumination. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, volume 20, pages 1025–1039, 1998.

[11] B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *IJCAI'81: Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[12] I. Matthews and S. Baker. Active Appearance Models Revisited. *International Journal of Computer Vision (IJCV)*, 60:135–164, 2004.

[13] I. Matthews, J. Xiao, and S. Baker. 2D vs. 3D Deformable Face Models: Representational Power, Construction, and Real-Time Fitting. *International Journal of Computer Vision (IJCV)*, 75(1):93–113, 2007.

[14] J. Peyras, A. Bartoli, H. Mercier, and P. Dalle. Segmented AAMs Improve Person-Independent Face Fitting. In *BMVC'07 - Proceedings of the 18th British Machine Vision Conference*, 2007.

[15] S. Romdhani. *Face Image Analysis using a Multiple Feature Fitting Strategy*. PhD thesis, University of Basel, Switzerland, 2005.

[16] S. Romdhani and T. Vetter. Efficient, Robust and Accurate Fitting of a 3D Morphable Model. In *ICCV'03: Proceedings of the 9th International Conference on Computer Vision*, volume 1, pages 59–66, 2003.

[17] P. H. Schoenemann and R. Carroll. Fitting One Matrix to Another Under Choice of a Central Dilation and a Rigid Motion. *Psychometrika*, 35(2):245–255, 1970.

[18] J. P. Thirion. Image matching as a diffusion process: An analogy with Maxwell's demons. *Medical Image Analysis*, 2:243–260, 1998.

[19] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid Structure-from-Motion: Estimating Shape and Motion with Hierarchical Priors. *TPAMI*, 30(5):878–892, 2008.