Learning a generic 3D face model from 2D image databases using incremental structure from motion

Jose Gonzalez-Mora[1,*], Fernando De la Torre[b], Nicolas Guil[1,*], Emilio L. Zapata[1]

[a]*Department of Computer Architecture, University of Malaga, Malaga, Spain 29071*
[b]*Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA 15213*

Abstract

Over the last decade 3D face models have been extensively used in many applications such as face recognition, facial animation and facial expression analysis. 3D Morphable Models (MMs) have become a popular tool to build and fit 3D face models to images. Critical to the success of MMs is the ability to learn a generic 3D face model. Towards that end, major limitations in the current MMs' process are: (1) collecting 3D data usually involves the use of expensive laser scans and complex capture setups, (2) the number of available 3D databases is limited, and typically there is a lack of expression variability, (3) finding correspondences and registering the 3D model is a labor intensive and error prone process.

This paper proposes an incremental Structure-from-Motion (SfM) approach to learn a generic 3D face model from large collections of existing 2D hand-labeled images, containing many subjects under different expressions and poses. Two major contributions are: (1) learning a generic 3D deformable face model from 2D databases, (2) incorporating a prior subspace into the incremental SfM formulation to provide robustness to noise, missing data and degenerate shape configurations. Experimental results on the CMU-PIE database show improvements in the generalization of the 3D face model across expression and identity.

Key words: Structure from Motion, incremental learning, Morphable Models, Active Appearance Models

1. Introduction

The face is a powerful channel of nonverbal communication and modeling faces has been useful in many computer vision applications such as virtual

*Corresponding author. Phone: (+34) 952133327. Fax: (+34) 952132790
Email addresses: jgmora@ac.uma.es (Jose Gonzalez-Mora), ftorre@cs.cmu.edu (Fernando De la Torre), nico@ac.uma.es (Nicolas Guil), ezapata@ac.uma.es (Emilio L. Zapata)

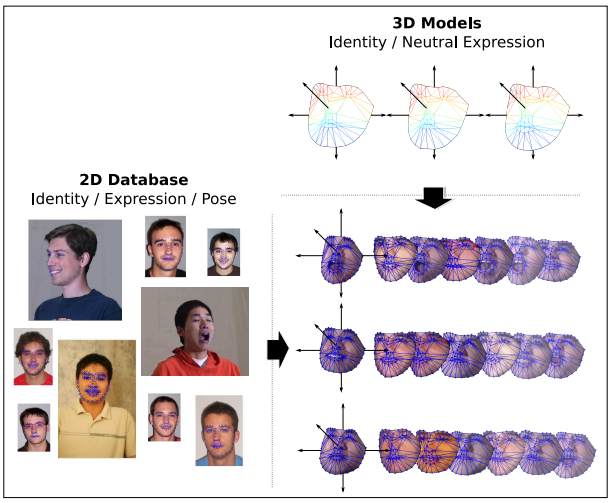

Figure 1: Learning a generic 3D face model. Left: 2D labeled images of several subjects with different expressions and poses. Top: A prior 3D model built from neutral expression shapes. Bottom right: Shape and texture of 3D faces with random expressions generated by the learned model

avatars [1], face recognition [2, 3], facial expression analysis [4], and model-based image coding [5]. Critical to the success of many computer vision face related applications is the ability to build a generic 3D face model that can generalize to untrained situations (e.g. different people, illumination, expressions). Moreover, the model should be able to decouple the facial deformations due to 3D motion and identity or expression.

Parameterized Appearance Models (PAMs) such as Eigentracking [6], Active Appearance Models (AMMs) [7, 8, 9, 10, 11, 12] and 3D Morphable models (3D-MMs) [13, 1, 14] have been popular tools to model the shape and appearance of faces in images. Although extensively used, PAMs have some limitations. One limitation of standard AAMs is its inability to decouple factors such as 3D motion, identity and expression. This is because a shape basis, modeled with Principal Component Analysis (PCA), is used to jointly model all these sources of variability. 3D Morphable Models [13, 1, 14] partially solve this problem by building 3D models that are able to decouple 3D pose changes from identity/expression. Although a promising approach, learning a generic 3D face model traditionally requires 3D face surface representations, which to date, has several drawbacks: (1) The number of available 3D databases is limited, and typically there is a lack of expression variability in the faces (e.g. standard databases include only neutral expressions). (2) Most common approaches use laser scans to get 3D models, requiring complex setups [15]. (3) Finding correspondences and registering the 3D model is a labor intensive and error prone process. Moreover, building generic 3D face models that generalize well across

several conditions requires large amounts of data and efficient algorithms should process the data incrementally.

Alternatively, over the last few years there has been a lot of interest in Structure-from-Motion (SfM) techniques to build 3D models from video. Previous work [16, 17, 18, 19, 20] built accurate person-specific face models from video sequences. However, there are several challenges when applying SfM techniques to the problem of learning a generic 3D face model from static 2D databases:

- Large baseline ranges between different images make it difficult to use SfM methods that have temporal smoothness assumptions between frames (e.g. [21]).
- Noise in landmarks due to the manual labeling process usually translates into degenerate SfM solutions (e.g. [22]).
- Self-occlusions cause certain landmarks to appear only in a small set of views resulting in large amounts of missing data [23, 24].
- Some face databases include only quasi-frontal images. Reduced parallax can generate low accuracy in the recovered depth.
- Small number of views per subject results in inaccurate 3D models. Several face databases contains few images per person (e.g. mugshot in police records).

To mitigate these problems this paper proposes an incremental algorithm for non-rigid SfM that builds a generic face model from a collection of 2D hand-labeled images. The main contributions are: (1) learning a generic 3D face model from 2D image databases, (2) an incremental approach to use prior 3D shape information in the SfM formulation. This prior information acts as a regularizing term reducing the spatial error in the recovered 3D structures. Furthermore, we provide an analysis of the generalization properties of the linear models generated by our SfM algorithm based on rank constraints. Figure 1 illustrates the main goal of this paper.

The rest of the paper is organized as follows. Section 2 reviews previous work on SfM techniques and 3D face model building techniques. Section 3 describes our incremental non-rigid SfM algorithm. Firstly, we present a short introduction to the non-rigid SfM problem, followed by a description of the main limitations of previous SfM approaches. Secondly, we describe the formulation of our algorithm. Finally, we propose a procedure to densify the model and extract the appearance model from the training images. Section 4 presents experimental results on building 3D face models from 2D image databases, analyzing the geometrical accuracy of the achieved reconstructions and evaluating the generalization properties to new individuals. Section 5 finalizes the paper with the conclusions and future work.

2. Previous work

This section describes previous work on non-rigid SfM and face modeling.

2.1. Non-rigid structure from motion

Structure-from-Motion (SfM) algorithms have been extensively used to factorize the rigid and non-rigid 3D structure of objects from a set of 2D point tracks. Early work by Tomasi and Kanade in the 90's [25] proposed a factorization approach to recover the shape of rigid objects from an orthographic camera. Over the past few years the factorization framework has been extended to deal with missing data [26, 27], more camera models (such as paraperspective or projective cameras), and non-rigid structure [16, 17, 18, 19, 20].

Bregler et al. [16] described a factorization method for objects with non-rigid structure where any 3D shape configuration is modeled as a linear combination of basic shapes defining principal deformation modes. Assuming a weak perspective camera projection, [16] proposed a factorization method that exploits rank constraints on camera rotations to recover non-rigid 3D shape and motion. Recently several authors [17, 18, 19] have shown that rotation constraints for the pose are not enough to achieve reliable 3D reconstructions. Brand [17] proposed an alternative optimization method by introducing extra constraints and forcing the deformation to be as small as possible (relative to the mean shape). Xiao et al. [18] proposed adding a set of constraints on the shape basis to recover better 3D models. These constraints are based on the assumption that there are n image frames (where n is the number of basis shapes) in which the basis shapes are known to be independent. However, as it was later pointed out by Brand [19], the algorithm breaks down with noisy data or when n is not correctly estimated.

Alternated least squares (ALS) [26, 28, 29] and expectation maximization (EM) [27, 21, 20] techniques have proven to be efficient methods to factorize the shape and motion components in SfM algorithms. These methods have been extended to incorporate missing data and to handle multiple view cases (where multiple projections of the same shape configuration are available). In presence of noisy data (e.g. inconsistencies in the tracked points or missing data), many SfM problems become ill-posed and Singular Value Decomposition (SVD) formulations are not effective. Torresani et al. [28] proposed a simple solution based on an alternated minimization scheme with promising results even when missing data is present. Buchanan and Fitzgibbon [23, 30] presented a class of second-order optimizations which converged more reliably than alternation approaches in different experiments. However, they concluded that for many real SfM problems it is not sufficient to minimize the reprojection error in order to get meaningful results and pointed out the need to further analyze the use of prior information. In many scenarios estimating deformable 3D shapes is inherently underconstrained, especially when using monocular 2D features, and standard SfM algorithms give degenerate solutions. Torresani et al. [21, 20] used an expectation maximization approach to solve the factorization problem, assuming Gaussian priors over the deformation parameters in order to avoid arbitrary variations. Del Bue et al. [31] enforced priors over the rigidity of some points to obtain reliable estimates of the object's rigid component. Olsen and Bartoli [32] imposed temporal smoothness and continuous variation in shape reconstructions. Similar in spirit to the approach presented in this paper, Del

Bue [33] introduced prior knowledge in the SfM algorithm in the form of previously known 3D shapes representing feasible configurations of the object, which at the end were used to regularize the rigid component of a deformable object. The formulation of this previous work is based on a factorization framework and prior information is incorporated into an intermediate affine solution but not the final metric reconstruction. Moreover, it is not clear how to incorporate missing data into the formulation. In this paper, we extend existing approaches by incorporating prior information for morphable shape models into the final Euclidean linear basis reconstruction and provide experimental results in difficult scenarios (e.g. severe occlusions and reduced number of views).

2.2. Modeling faces from images

Parameterized Appearance Models (PAMs) such as Eigentracking [6], Active Appearance Models (AMMs) [7, 8, 9, 10, 11, 12] and 3D Morphable models (3DMMs) [13, 1, 14] have been a popular tool to model the shape and appearance of faces in images. Using labeled data, PAMs learn a shape and appearance model by computing PCA on a set of landmarks and registered textures after a previous normalization with Procrustes analysis. Recently, [34] pointed out how registration and modeling of deformable bases are coupled problems. By representing the identity/expression changes as a random noise in the Procrustes procedure, the resulting bases are biased. Alternatively, SfM algorithms can be used to jointly optimize motion and a non-rigid 3D basis. Previous work [16, 17, 18, 19, 20] has shown how to build accurate person-specific models from video, where depth information can be extracted from a sequence containing wide range of poses and where it is possible to apply motion continuity constraints. However, it is unclear how to extend these approaches to build generic face models from existing static image databases, especially those containing a reduced set of views (i.e frontal galleries of common biometric databases).

Xiao et al. [35] showed how a non-rigid 3D model is built using SfM combined with a 2D-AAM facial feature tracker. The 3D shape model is used as a regularization term in the 2D tracking system, resulting in the "Combined 2D+3D AAM" algorithm. Similarly to our method, the model is built from a set of manually labeled static images, however in [35] the authors learned a person-specific shape basis for each tracking sequence. This basis does not intend to be a generic face model, and consequently, no evaluation of the geometrical accuracy, 3D pose disambiguation or generalization properties of the SfM algorithm is presented. Also, the SVD-based factorization approach used in [35] makes it difficult to deal with a wide range of rotation angles that would cause occlusions and missing data.

Commonly used face image databases are labeled using a small set of landmarks in order to reduce the manual intervention. As a result, models generated from this training data have lower resolution than those obtained from dense scans [36]. Consequently, they may reproduce unrealistic projections for certain views or fail to accurately model the surface normals (e.g., when computing light incidence angles). A possibility to obtain dense shape descriptions without requiring exhaustive laser captures is to represent novel face configurations

by morphing neutral expression scans into a desired expression/identity. Several techniques have been described to learn 3D shape deformations modeling expression and identity changes in human faces. Some of them [37, 38] require a small control group of subjects for which all the possible expressions have been captured using laser scanners. Other approaches define the target 3D shape by means of a SfM algorithm, eliminating the need of extra captures. Kim et al. [39] use a factorization-based rigid SfM algorithm on the AAM point tracks of a specific video sequence representing the target facial configuration.

In the following section we propose a novel 3D Morphable Model building algorithm based on an incremental non-rigid SfM technique, which is able to extract 3D surface descriptions from a collection of hand-labeled static images containing multiple individuals and expressions.

3. Non rigid Structure from Motion

This section describes our approach to incremental SfM. Subsection 3.1 introduces the SfM problem and ALS techniques. Subsection 3.2 describes the main limitations of previous SfM approaches while subsections 3.3, 3.4 and 3.5 present our formulation.

3.1. Problem formulation

A common approach [16] to model the non-rigid 3D structure of deformable objects is to use a linear combination of the deformation basis vectors $\mathbf{S} = \{\mathbf{s}_l \in \Re^{3d\times 1}\}, l = 1, ..., n$, where d denotes the number of points, and n the number of basis components. A 3D shape instance, $\mathbf{s}$, can be generated using a linear combination of these basis vectors:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{l=1}^{n} c_l \mathbf{s}_l = \bar{\mathbf{s}} + \mathbf{S}\mathbf{c} \tag{1}$$

where $\mathbf{c} \in \Re^n$ are the coefficients of the linear combination.

Using a RTS (rotation-translation-scale) camera model we can express the projection of the shape onto image plane $\mathbf{p}$ as:

$$\mathbf{p} = \begin{bmatrix} x \\ y \end{bmatrix} = k\mathbf{R}\mathbf{s} + \mathbf{t} \tag{2}$$

where k is a scale parameter, $\mathbf{R} \in \Re^{2\times 3}$ is a matrix containing the first two rows of the 3D rotation matrix, and $\mathbf{t} = (t_x, t_y)^T$ is the translation. More accurate perspective models exist, but this approximation has previously been effective in similar contexts [25, 16] (especially if the face is far from the camera).

Let us consider now that we have q of these 2D shape projections, $\{\mathbf{p}_i\}, i = 1 \cdots q$, representing several instances of a deformable object class projected with pose parameters k_i, $\mathbf{R}_i$ and $\mathbf{t}_i$. To solve the SfM problem, we minimize the following objective function:

$$\min_{\mathbf{S},\mathbf{c}_i,k_i,\mathbf{R}_i,\mathbf{t}_i} \sum_{i=1}^{q} \|\mathbf{p}_i - k_i\mathbf{R}_i\left(\bar{\mathbf{s}} + \mathbf{S}\mathbf{c}_i\right) - \mathbf{t}_i\|_2^2 \tag{3}$$

A commonly used technique when there is no missing data on $\mathbf{p}_i$ is SVD factorization [16, 18, 19]. Alternatively, the previous equation can be solved using an Alternating Least Squares (ALS) technique [40, 20, 29].

The first step of the ALS optimization solves for the pose and shape coefficients (equation 3) given the current estimate of shape deformation basis $\mathbf{S}$, and the mean shape $\bar{\mathbf{s}}$. Because of the non-linear dependency of the shape projections $\mathbf{p}_i$ for the pose parameters (in particular, the rotation angles), this step is also an iterative procedure. For every shape projection $\mathbf{p}_i$, we successively compute refined shape coefficients and pose increments (Δk_i, $\Delta \mathbf{R}_i$, $\Delta \mathbf{t}_i$) which are then used to update scale, rotation and displacement estimates (k_i, $\mathbf{R}_i$ and $\mathbf{t}_i$).

The second step upgrades the mean shape $\bar{\mathbf{s}}$ and the shape deformation basis $\mathbf{S}$. Considering that occlusions, due to pose changes, can introduce missing data, we separately estimate each component j of $\bar{\mathbf{s}}$ and $\mathbf{S}$ using the set $1:I$ of shape projections for which the point j is visible, by means of the following equation:

$$vec(\hat{\mathbf{S}}^{(j)}) = \mathbf{M}^{+}(\mathbf{p}_{1:I}^{(j)} - \mathbf{t}_{1:I}) \tag{4}$$

where $\mathbf{M}^{+}$ denotes the pseudoinverse of $\mathbf{M}$, and:

- $\hat{\mathbf{S}} = [\bar{\mathbf{s}}, \mathbf{S}]$, $\hat{\mathbf{c}} = [1, \mathbf{c}^T]^T$ is the concatenation matrix of the mean shape and the deformation basis.
- $\mathbf{p}_{1:I}^{(j)} = [{\mathbf{p}_1^{(j)}}^T, \cdots, {\mathbf{p}_I^{(j)}}^T]$ is the projection vector containing all visible views of point j
- $\mathbf{t}_{1:I} = [\mathbf{t}_1^T, \cdots, \mathbf{t}_I^T]^T$ is the translation vector containing the corresponding displacements for each view
- $\mathbf{M} = [\hat{\mathbf{c}}_1 \otimes (k_1 \mathbf{R}_1^T), \cdots, \hat{\mathbf{c}}_I \otimes (k_I \mathbf{R}_I^T)]$ combines shape parameters and motion to describe the shape configuration for each view.

3.2. Limitations for the Alternated Least Squares Algorithm

The ALS algorithm has two major drawbacks. Firstly, although it effectively minimizes the reprojection error, the least squares solution for this objective function does not necessarily correspond to a realistic solution [22]. For instance, figure 2 represents the recovered 3D shape by applying rigid SfM to a given subject using 3 views. It is a very simple example illustrating the problem of degenerated solutions in SfM algorithms (real world applications will usually involve more extensive datasets). Image *(a)* shows the results using an ALS optimization. The reconstruction is unrealistic even when the reprojection error is low. Image *(a)* shows a reconstructed mesh with overscaled Z component and wrong camera orientations that, although geometrically feasible, is far from the real solution. The real scene configuration is represented in image *(b)*.

Secondly, ALS approaches are prone to local minima and are very sensitive to initial values. As a simple illustrative example, figure 3 shows the results of an ALS algorithm used to build a non-rigid model representing 3 expressions for

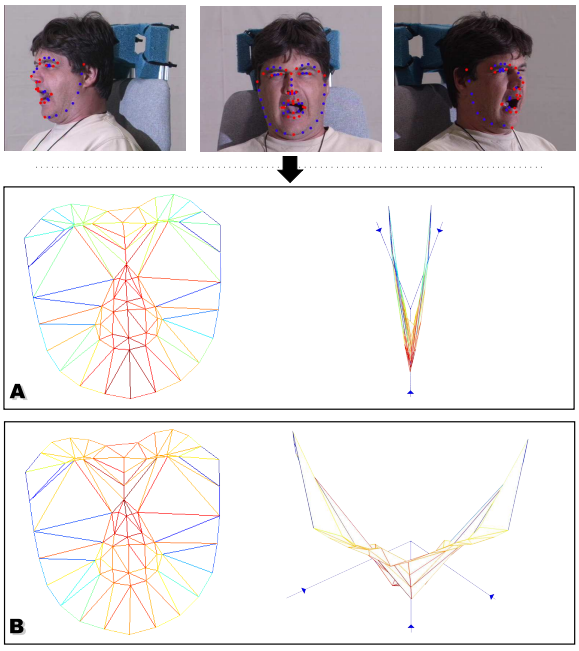

Figure 2: Example of rigid 3D structure recovering. The images were captured using approximately -60, 0 and 45 degrees jaw rotation. *A*: mesh obtained using an alternated least squares algorithm without prior information. *B*: ground truth mesh.

a given subject (neutral, smile and scream) with different initializations. Figure 4 represents the evolution of the reprojection error in each iteration of the algorithm (50 iterations). Similarly to [21], method *(a)* initializes the algorithm with the SVD of the observation matrix containing all shape projections. This initialization value is far away from the correct solution, and the algorithm results in an unrealistic model. One of the main factors that justifies the poor results achieved with this initialization method is the severe presence of missing data, due to pose occlusions which yields a bad estimation of the mean shape. Image *(b)* shows an alternative initialization method based on [41]. It fits a low rank matrix to the projections, filling the missing elements, for which the SVD can obtain a better mean shape estimation. Although the recovered mesh is closer to the real shape and the obtained reprojection error is lower than previous algorithm, the results are not satisfactory. In figure *(c)* we use a previously existing set of 3D faces with neutral expression to initialize the algorithm. A initial value for the linear shape basis is obtained by applying PCA analysis to this 3D neutral dataset. Since this estimate is closer to a real solution, the reprojection error quickly decreases to a lower value and the solution is visually correct. Finally image *(d)* shows a feasible solution obtained by our algorithm, using 3D neutral expression face shapes as a prior as it will be described in next sections. It is interesting to note that the reprojection error in this case is bigger than some of the previous approaches even when the solution is more realistic. This is caused by the regularizing component introduced by the prior.

3.3. Joint pose parameters estimation

Unlike previous approaches [28, 20], we compute a joint estimation of the scale, rotation and translation parameters. In particular, we build a *motion basis* representing a linear approximation of all possible changes in 2D shape due to pose changes around the current estimate of pose (tangent space). This motion basis determines a gradient descent direction in which all pose parameters are jointly optimized, preventing local minima solutions yielded by the consecutive estimation.

Given an initial estimate of the shape parameters $\mathbf{c}$ in equation 3, we minimize the shape reconstruction error with respect to pose parameters by determining an optimal increment for the rotation $\Delta\mathbf{R}$, scale Δk and displacement $\Delta\mathbf{t}$. The parameter increments are computed using a linear optimization scheme successfully used in previous techniques for robust model fitting (i.e. [42]).

$$\mathbf{p} = k\Delta k\mathbf{R}\Delta\mathbf{R}\left(\bar{\mathbf{s}} + \sum_{i=1}^{n} c_i\mathbf{s}_i + \mathbf{t} + \Delta\mathbf{t}\right) \tag{5}$$

For small rotation angles $\alpha, \beta, \gamma \ll 1$, the rotation matrix $\Delta\mathbf{R}$ can be approximated as:

$$\Delta\mathbf{R} = \Delta\mathbf{R}_\alpha\Delta\mathbf{R}_\beta\Delta\mathbf{R}_\gamma \approx \begin{bmatrix} 1 & -sin\alpha & sin\gamma \\ sin\alpha & 1 & -sin\beta \\ -sin\gamma & sin\beta & 1 \end{bmatrix} \tag{6}$$

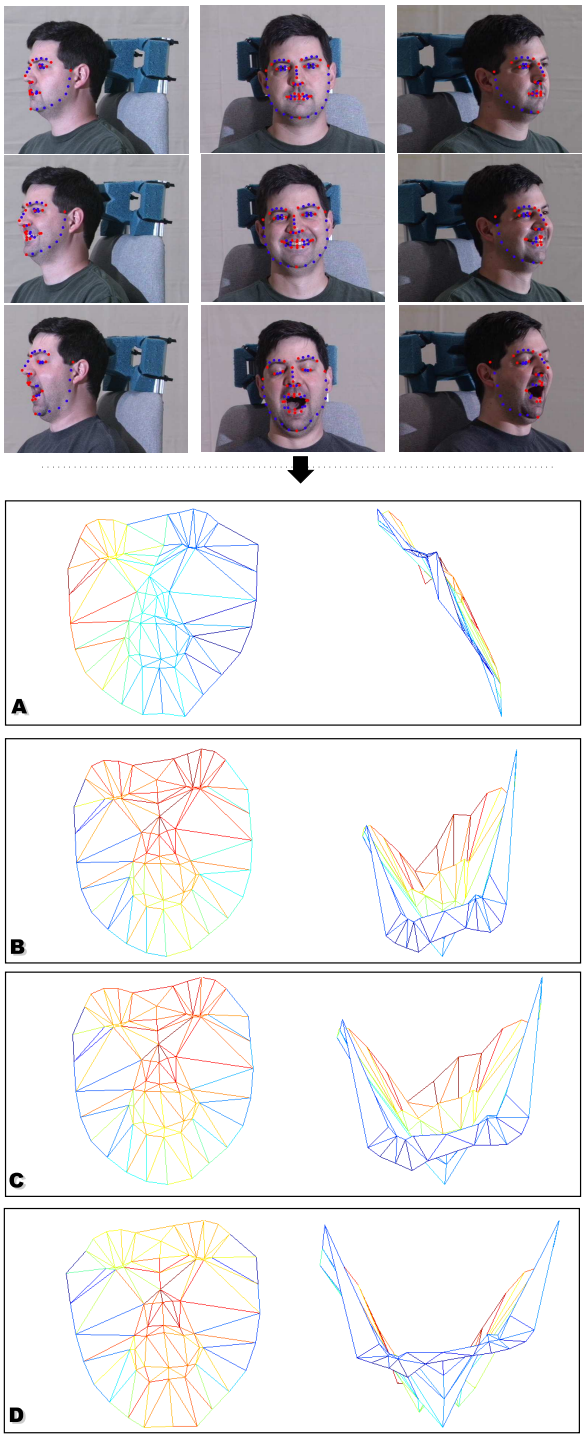

Figure 3: Non-Rigid 3D structure recovering example. Several initializations are used: *(a)* initializing the algorithm with a rigid mesh obtained using SVD, *(b)*: filling-in the missing data in the observation matrix, *(c)*: using a neutral expression basis, *(d)*: incremental SfM

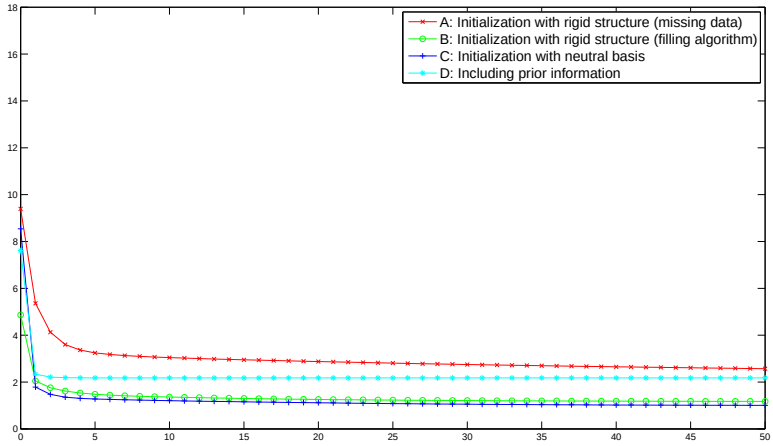

Figure 4: Reprojection error for different initializations of an ALS algorithm

Using this approximation, a rotated point $\mathbf{R}\,[x,y,z]^T$ can be expressed as:

$$\Delta\mathbf{R}\begin{bmatrix} x \\ y \\ z \end{bmatrix} \approx \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \tag{7}$$
$$+sin\alpha\begin{bmatrix} -y \\ x \\ 0 \end{bmatrix} + sin\beta\begin{bmatrix} 0 \\ -z \\ y \end{bmatrix} + sin\gamma\begin{bmatrix} z \\ 0 \\ -x \end{bmatrix}$$

Applying this approximation to every shape point, and ignoring the effects of the rotation in the shape deformation vectors and translation, we have that shape rotation can be approximated in the tangent space using an extra linear basis including vectors $\mathbf{s}_\alpha$, $\mathbf{s}_\beta$, $\mathbf{s}_\gamma$:

$$\Delta\mathbf{R}\left(\bar{\mathbf{s}} + \sum_{l=1}^{n} c_l\mathbf{s}_l + \mathbf{t}\right) \approx \tag{8}$$
$$\bar{\mathbf{s}} + \sum_{l=1}^{n} c_l\mathbf{s}_l + \mathbf{t} + \Delta\mathbf{t} + c_\alpha\mathbf{s}_\alpha + c_\beta\mathbf{s}_\beta + c_\gamma\mathbf{s}_\gamma$$

with coefficients $c_\alpha = sin\alpha$, $c_\beta = sin\beta$ and $c_\gamma = sin\gamma$ and

$$\begin{aligned} \mathbf{s}_\alpha &= [-\bar{y}_1, \bar{x}_1, 0, -\bar{y}_2, \bar{x}_2, 0, \cdots]^T \\ \mathbf{s}_\beta &= [0, -\bar{z}_1, \bar{y}_1, 0, -\bar{z}_2, \bar{y}_2, \cdots]^T \\ \mathbf{s}_\gamma &= [\bar{z}_1, 0, -\bar{x}_1, \bar{z}_2, 0, -\bar{x}_2, \cdots]^T \end{aligned} \tag{9}$$

Using this approximation, equation 5 can be written as:

$$\mathbf{p} = k\Delta k\mathbf{R}\left(\bar{\mathbf{s}} + \mathbf{S}\mathbf{c} + \tilde{\mathbf{S}}\tilde{\mathbf{c}} + \mathbf{t}\right) \tag{10}$$

where:

- $\mathbf{S}$ and $\mathbf{c}$ are the initial shape linear basis and coefficients.
- $\tilde{\mathbf{S}} = [\mathbf{s}_\alpha \mathbf{s}_\beta \mathbf{s}_\gamma \mathbf{s}_x \mathbf{s}_y \mathbf{s}_z]$ and $\tilde{\mathbf{c}} = [c_\alpha c_\beta c_\gamma \Delta t_x \Delta t_y \Delta t_z]^T$ are the motion linear basis, with $\mathbf{s}_x = [1\ 1 \cdots 0\ 0 \cdots 0\ 0]^T$, $\mathbf{s}_y = [0\ 0 \cdots 1\ 1 \cdots 0\ 0]^T$ and $\mathbf{s}_z = [0\ 0 \cdots 0\ 0 \cdots 1\ 1]^T$ displacement vectors of $3 \times p$ elements.

Subtracting $k\mathbf{R}\bar{\mathbf{s}}$ in each term

$$\begin{aligned}\mathbf{p} - k\mathbf{R}\bar{\mathbf{s}} = k\mathbf{R}\bar{\mathbf{s}}(\Delta k - 1) + \\ k\Delta k\mathbf{R}\left(\mathbf{S}\mathbf{c} + \tilde{\mathbf{S}}\tilde{\mathbf{c}} + \mathbf{t}\right)\end{aligned} \tag{11}$$

Considering $k\Delta k \approx k$ in the second term and introducing the error term $\mathbf{e} = \mathbf{p} - k\mathbf{R}\left(\bar{\mathbf{s}} + \mathbf{t}\right)$, we get a linear relationship between current shape errors and parameters increment that can be used to refine the estimation through an iterative algorithm:

$$\mathbf{e} = k\mathbf{R}\mathbf{S}'\mathbf{c}' \tag{12}$$

where $\mathbf{S}'$ and $\mathbf{c}'$ are the extended shape and motion basis and extended parameters vector, including $n + 7$ shape vectors and coefficients respectively:

$$\mathbf{S}' = \left[\mathbf{S}, \bar{\mathbf{S}}\right] = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n, \bar{\mathbf{s}}\mathbf{s}_\alpha, \mathbf{s}_\beta, \mathbf{s}_\gamma, \mathbf{s}_x, \mathbf{s}_y, \mathbf{s}_z] \tag{13}$$

$$\mathbf{c}' = [\mathbf{c}, \bar{\mathbf{c}}] = [c_1, c_2, \cdots, c_n, (\Delta k - 1), c_\alpha, c_\beta, c_\gamma, t_x, t_y, t_z]^T \tag{14}$$

Similarly to [23], the use of a non-linear optimization to estimate the pose parameters instead of alternately computing the scale, rotation and translation avoids "flatlining" problems and yields accurate solutions. When used in our SfM algorithm, we force the shape deformation basis to be orthogonal to this explicitly modeled motion basis and it will result in better decoupling of the structure and motion components, producing more accurate reconstructions as shown in the experimental results.

3.4. Multiple views

The accuracy of ALS techniques can be improved by incorporating multiple training views of the same individual in the formulation. In the multi-view version of the ALS technique we modify the first alternating step to fit the shape basis in multiple views of the same shape configuration, obtaining the pose and shape parameters. Simultaneously, we adapt the second alternating step that computes the shape basis by simply adding extra rows on both sides of equation 4 including the extra views provided for the different shape instances. Considering that each point j is seen from two views for a set of I shape configurations,

then matrix $\mathbf{M}$, projections vector $\mathbf{p}_{1:I}$ and translation vector $\mathbf{t}_{1:I}$ would have the following form:

$$\mathbf{p}_{1:I}^{(j)} = [{\mathbf{p}_{1,1}^{(j)}}^T, {\mathbf{p}_{1,2}^{(j)}}^T, \cdots, {\mathbf{p}_{I,1}^{(j)}}^T, {\mathbf{p}_{I,2}^{(j)}}^T] \tag{15}$$

$$M = [\hat{\mathbf{c}}_1(k_{1,1}\mathbf{R}_{1,1}^T), \hat{\mathbf{c}}_1(k_{1,2}\mathbf{R}_{1,2}^T), \cdots, \hat{\mathbf{c}}_I(k_{I,1}\mathbf{R}_{I,1}^T), \hat{\mathbf{c}}_I(k_{I,2}\mathbf{R}_{I,2}^T)]^T \tag{16}$$

$$\mathbf{t}_{1:I} = [\mathbf{t}_{1,1}^T, \mathbf{t}_{1,2}^T, \cdots, \mathbf{t}_{I,1}^T, \mathbf{t}_{I,2}^T]^T \tag{17}$$

where $\mathbf{p}_{i,v}$, $k_{i,v}$, $\mathbf{R}_{i,v}$ and $\mathbf{t}_{i,v}$ denote the projection view v for a shape i.

3.5. Adding prior shape information

Incremental learning (e.g. [43, 44]) plays an essential role in many subspace based methods to solve computer vision problems. In the SfM problem domain, we show how it constitutes an efficient approach to integrate information from multiple sources (3D shape descriptions and 2D manually labeled images) in order to create accurate 3D face models using a computationally efficient approach.

To prevent local minima and provide robustness to noise, our algorithm regularizes the SfM by incorporating previous knowledge about possible solutions. Similarly to [33], we incorporate prior information in the form of a set $\mathbf{\Pi}_0$ of m 3D shapes that are already known to be feasible configurations for the considered deformable object.

$$\mathbf{\Pi}_0 = \left[\mathbf{s}_0^1, \mathbf{s}_0^2, \cdots, \mathbf{s}_0^m\right] \tag{18}$$

We compute a prior shape basis $\mathbf{S}_0$ by applying Principal Component Analysis (PCA) on these known shapes. It is used at the first iteration to initialize the shape deformation basis estimate. Moreover, $\mathbf{S}_0$ is combined with the information extracted from training data at successive iterations to regularize the solution and force feasible shape basis estimations. Unlike previous algorithms [33], priors are directly added to the final Euclidean description of the shape basis without the need to compute a metric upgrade in a separate step. In the following, we apply a SVD updating technique [45, 43] to formulate an incremental shape basis update step that makes use of the available prior information.

As mentioned in section 3.1, the first step of the ALS-SfM algorithm obtains a set of shape parameters $\mathbf{c}_i$. These parameters will generate approximated shape instances $\mathbf{s}_i$ of the object using the current estimate of the shape basis. In order to get a new shape basis update minimizing the reprojection error, we compute a set of 3D shape corrections $\Delta\mathbf{s}_i$ to refine these shape estimates ($\mathbf{s}_i \leftarrow \mathbf{s}_i + \Delta\mathbf{s}_i$). We deal with missing data computing the shape corrections for each point j separately, using the input shape projections and the obtained values for pose parameters:

$$\mathbf{p}_i^{(j)} - k\mathbf{R}(\bar{\mathbf{s}}^{(j)} + \mathbf{S}^{(j)}\mathbf{c}_i) - \mathbf{t} = k\mathbf{R}\Delta\mathbf{s}_i^{(j)} \tag{19}$$

$$\Delta \mathbf{s}_i^{(j)} = (k\mathbf{R})^+ \left(\mathbf{p}_i^{(j)} - k\mathbf{R}(\bar{\mathbf{s}}^{(j)} + \mathbf{S}^{(j)}\mathbf{c}) - \mathbf{t} \right) \tag{20}$$

Let us denote by $\mathbf{\Pi}$ the 3D shape estimates matrix obtained by concatenation of the q updated shapes $\mathbf{s}_i$.

$$\mathbf{\Pi} = [\mathbf{s}_1, \cdots, \mathbf{s}_q] \tag{21}$$

Now, we determine the updated shape basis that best represents (for a given rank) the 3D shape estimates $\mathbf{\Pi}$ (obtained from the provided shape projections $\mathbf{p}$) and the subspace spanned by the provided prior basis $\mathbf{S}_0$ (with associated eigenvalues λ_0). First, we update the mean shape estimate $\bar{\mathbf{s}}$:

$$\bar{\mathbf{s}} \leftarrow (\bar{\mathbf{s}} + \frac{\sum_i (\mathbf{s}_i)}{q})/2 \tag{22}$$

The coefficients of the updated shape collection in the prior basis will be given by:

$$\mathbf{c}_0 = \mathbf{S}_0^T \cdot \mathbf{\Pi} \tag{23}$$

Therefore, the component of $\mathbf{\Pi}$ orthogonal to the prior basis can be expressed as:

$$\mathbf{\Pi}_\perp = \mathbf{\Pi} - \mathbf{S}_0 \cdot \mathbf{c}_0 \tag{24}$$

We compute an orthogonal basis $\mathbf{S}_\perp$ spanning the subspace defined by $\mathbf{\Pi}_\perp$:

$$\mathbf{S}_\perp = qr(\mathbf{\Pi}_\perp) \tag{25}$$

and the coefficients of $\mathbf{\Pi}_\perp$ w.r.t this basis

$$\mathbf{c}_\perp = \mathbf{S}_\perp^T \cdot \mathbf{\Pi}_\perp \tag{26}$$

Let us define $\mathbf{S}_*$ and $\mathbf{\Lambda}$ as:

$$\mathbf{S}_* = [\mathbf{S}_0, \mathbf{S}_\perp] \tag{27}$$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_0 & \mathbf{c}_0 \\ \mathbf{0} & \mathbf{c}_\perp \end{bmatrix} \tag{28}$$

Applying SVD on the matrix $\mathbf{\Lambda}$:

$$[\mathbf{U}, \mathbf{D}, \mathbf{V}] = svd(\mathbf{\Lambda}) \tag{29}$$

the new shape vectors and their associate eigenvalues will be given by:

$$\mathbf{S} \leftarrow \mathbf{S}_* \cdot \mathbf{U} \tag{30}$$

$$\lambda \leftarrow diag(\mathbf{D}) \tag{31}$$

Considering the new associated right singular vectors matrix V' to be:

$$\mathbf{V}' = \begin{bmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{I} \end{bmatrix} \mathbf{V} \tag{32}$$

the values of $\mathbf{S}$, λ and V' can be substituted to verify that they indeed correspond to the SVD of the joint observation matrix $[\mathbf{\Pi}_0\ \mathbf{\Pi}]$:

$$\begin{aligned} \mathbf{S}\lambda\mathbf{V'}^T = \mathbf{S}_*\mathbf{U}\mathbf{D}\mathbf{V}^T \begin{bmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{I} \end{bmatrix}^T = \mathbf{S}_*\mathbf{\Lambda} \begin{bmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{I} \end{bmatrix}^T = \\ = [\mathbf{S}_0\ \mathbf{S}_\perp] \begin{bmatrix} \lambda_0 & \mathbf{c}_0 \\ 0 & \mathbf{c}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{I} \end{bmatrix}^T = \\ = \begin{bmatrix} \mathbf{S}_0\ (\mathbf{I} - \mathbf{S}_0\mathbf{S}_0{}^T)\mathbf{\Pi}/\mathbf{c}_\perp \end{bmatrix} \begin{bmatrix} \lambda_0 & \mathbf{S}_0{}^T\mathbf{\Pi} \\ 0 & \mathbf{c}_\perp \end{bmatrix} \begin{bmatrix} \mathbf{V}_0 & 0 \\ 0 & \mathbf{I} \end{bmatrix}^T = \\ = \begin{bmatrix} \mathbf{S}_0\lambda_0\mathbf{V}_0{}^T\ \mathbf{\Pi} \end{bmatrix} = [\mathbf{\Pi}_0\ \mathbf{\Pi}] \end{aligned} \tag{33}$$

In this expression the SVD of the prior shape matrix is considered to be $\mathbf{\Pi}_0 = \mathbf{S}_0\lambda_0\mathbf{V}_0$). We finally limit the number of components in $\mathbf{S}$ and λ to the given rank in the updated basis.

4. Building dense 3D surface descriptions and appearance modeling

The resulting deformable shape model obtained with the SfM algorithm contains a reduced number of points corresponding to the selected landmarks in the images. In the following, we describe how to densify the model using Thin Plate Spline (TPS) as the mapping and interpolation tool for deformation transfer and synthesis.

TPS [46] is an effective tool for modeling coordinate transformations that has been successfully applied in several computer vision applications. It is a commonly used technique to represent coordinate mappings from $\Re^2$ to $\Re^2$. Let $\mathbf{x}_i$ denote the target function in locations $\mathbf{u} = (\mathbf{u_i}, \mathbf{v_i})^\mathbf{T}$ in the plane, with $i = 1, 2, \cdots, p$. We assume that the locations $\mathbf{u}$ are all different and are not collinear. TPS defines a interpolant function $f(\mathbf{u})$ that minimizes the bending energy and is formulated as:

$$f(\mathbf{u}) = \mathbf{c} + \mathbf{A}\mathbf{u} + \mathbf{W}^T s(\mathbf{u}) \tag{34}$$

where:

- $\mathbf{c}$, $\mathbf{A}$ and $\mathbf{W}$ are the TPS parameters.
- $s(\mathbf{u}) = (\sigma(\mathbf{u} - \mathbf{u}_1), \sigma(\mathbf{u} - \mathbf{u}_2), \cdots, \sigma(\mathbf{u} - \mathbf{u}_m))^T$, with $\sigma(\mathbf{r}) = \|\mathbf{r}\|$.

The deformation transfer problem is defined as follows: given a pair of point sets $\mathbf{x}_i = (x_i, y_i, z_i)^T$ and $\mathbf{x}'_i = (x'_i, y'_i, z'_i)^T$ with known correspondences on two surfaces, and considering $\mathbf{u}_i$ to be the common texture coordinate associated to

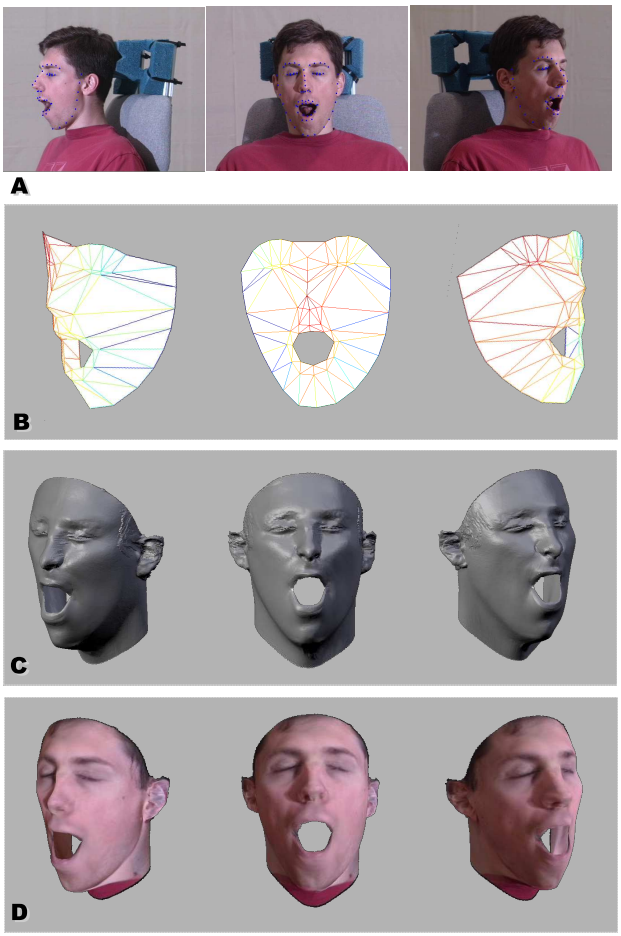

Figure 5: Building a 3D Morphable Model. (a) Images used in the training process for one subject (b) Different views of the sparse 3D geometrical reconstruction obtained from the SfM algorithm (c) Model after mesh densifying (d) Final model after texture mapping.

each point i, we can fit a TPS over $(u_i, v_i, x'_i - x_i)$ to get an interpolation deformation model for translation in x direction (similar for y and z) coordinates.

The 3D shape created using the SfM algorithm is projected into the prior dense shape basis to obtain a dense representation for each subject in the training set. This projection is not able to represent deformations not included in the prior database, such as novel expressions. Thus, we learn a deformation transfer interpolant using TPS mapping 3D locations of the landmark points in the achieved projection into the actual position obtained from the SfM algorithm. Using this deformation model we estimate the position of all the points in the dense model.

After having an accurate 3D shape description for each of the models in the training samples, the next step is to compute the appearance model. We construct a view independent cylindrical texture by capturing the pixel intensities from the training images in the areas situated under the shape projections. In order to avoid texture artifacts caused by the influence of pixels positioned in border areas, we compute a blending weight for each triangle on the face mesh for each image based on the angle between surface normal and the camera direction. A test of visibility is performed using a z-buffer method [47]. When the triangle is invisible, its weight is set to 0.0 and all the weights are then normalized so that the sum over all the images is equal to 1.0. Figure 6 represents the blending weight maps for three different views of a face and the final texture obtained by fusing the weighted pixel intensities captured from these views. Finally, we apply PCA to the set of textures obtained for the different face instances to build a linear appearance model.

5. Experimental results

This section describes experimental results evaluating the performance of our incremental SfM algorithm and comparing it with other approaches. These experiments use labeled facial images from the CMU Multi-PIE database [48].

The CMU Multi-PIE has 337 subjects simultaneously recorded with 15 cameras and 6 expressions. It has been partially labeled using 68 distinctive shape landmarks. These landmarks were manually located by trained personnel using three basic rules:

- "Non-ambiguous" landmarks (those whose position can be easily determined based on the facial structure, i.e. nose tip) must be consistently placed across different individuals and expressions. I.e. noise tip always correspond to the minimum Z coordinate.

- Those landmarks for which is hard to define a exact location (i.e. points along the jaw contour) are equispaced between the "non-ambiguous" landmarks.

- Not all the landmarks are visible in every view due to face self occlusions. For each pose we only determine the position of a subset of points that

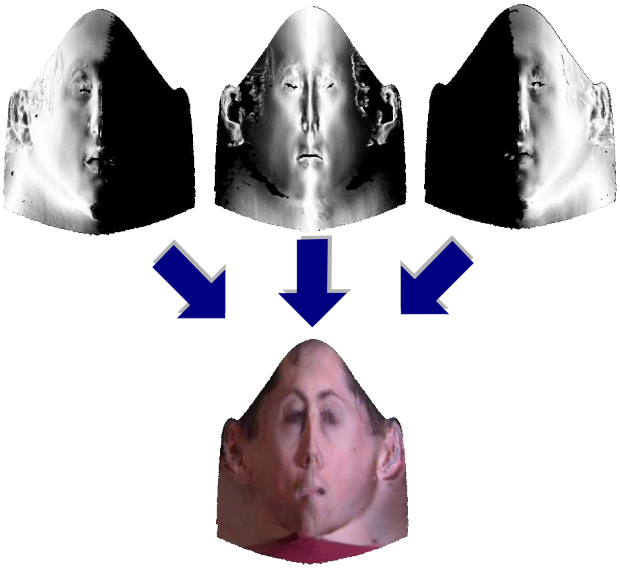

Figure 6: Texture mapping process. Different weight maps are obtained from the given views to combine the pixel intensities into a final texture map

Figure 7: Sample labeled image

are not occluded (i.e. we consider 39 landmarks for -60 and 45 degrees rotations)

In despite of these rules, the low texture information in many areas of human faces makes it hard to accurately match landmark locations between different subjects, expressions and poses even for non-ambiguous landmarks. It results in a noisy landmark dataset. See figure 7 for an example of the labeled images (points considered "non-ambiguous" are represented in red, the remaining ones in blue).

Given that no 3D information is originally provided with the CMU Multi-PIE database, we initially built 3D ground-truth shape models for a subset of testing subjects using SfM. We used the incremental SfM algorithm proposed in this paper to extract these reference shapes but, contrarily to the models generated during the experiments, the ground-truth estimation was performed using 13 views and manual correction of the reprojection errors to ensure the accuracy of the face reconstructions. In the experiments, the models are built using 1, 2 o 3 views, a more realistic training set for most common face databases.

Additionally, we incorporate the prior 3D information required by our incremental SfM algorithm from a set of 16 tridimensional faces with neutral expression. These shapes were generated as instances of a Morphable Model built from laser scans [13]. The dimension of the shape representations in the prior is reduced by matching the considered 68 face landmarks with corresponding 3D points in the dense Morphable Model.

5.1. Face model construction using incremental SfM

This section evaluates the geometrical reconstruction error of our incremental non-rigid SfM algorithm. We select 30 subjects representing "neutral", "smile" and "scream" expressions from the Multi-PIE database. The SfM algorithm builds a non-rigid 3D model for each expression that is used to represent the different shape instances corresponding to different subjects. The geometrical reconstruction error for every individual with respect to 3D ground truth shapes is represented in figures 8, 9 and 10. This geometrical error is defined as the mean 3D Euclidean distance between the points in the ground truth shape and the corresponding points in the obtained shape reconstruction (after both shapes are aligned removing rigid transformations). The experiments are repeated 3 times using one view per person (a frontal view, 0^o of jaw rotation), two views (0^o and 45^0) or three views (0^o, 45^o and -60^o).

We compare the results of our incremental SfM algorithm with an ALS optimization algorithm ([28] or BCD-LS algorithm in [20]) that shows competitive performance when compared with state of the art techniques [20]. Two variants are tested for the incremental SfM: (a) one using a consecutive least-squares estimation of scale, rotation and translation in the pose estimation step (similar to the original presented in [28, 20]), and (b) using a joint estimation of the pose parameters (section 3.3). We empirically chose the rank of the learned models for the different expressions to be 6 as it provides the more accurate reconstructions.

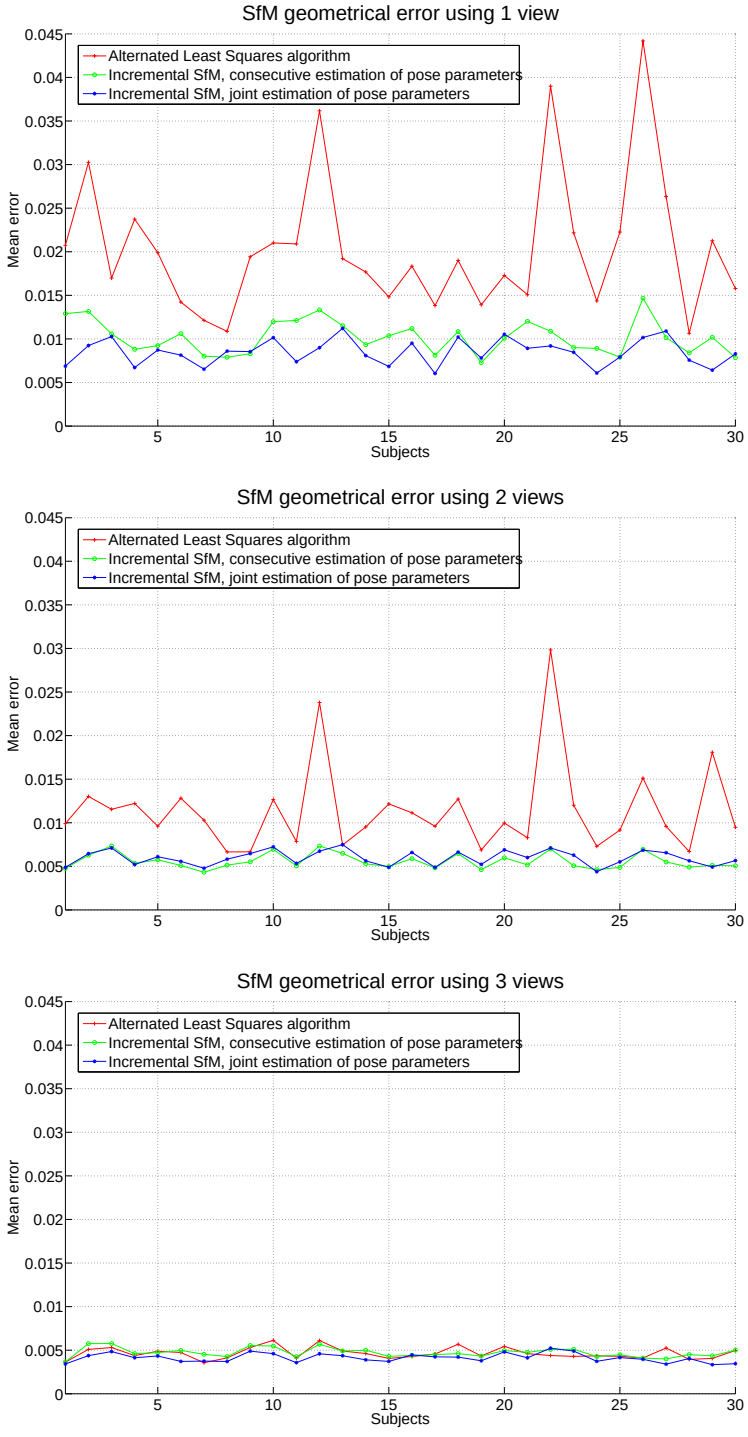

Figure 8: Geometrical error of the recovered 3D shapes for different individuals (**neutral expression**), using 1 view (*top*), 2 views (*middle*) and 3 views (*bottom*)

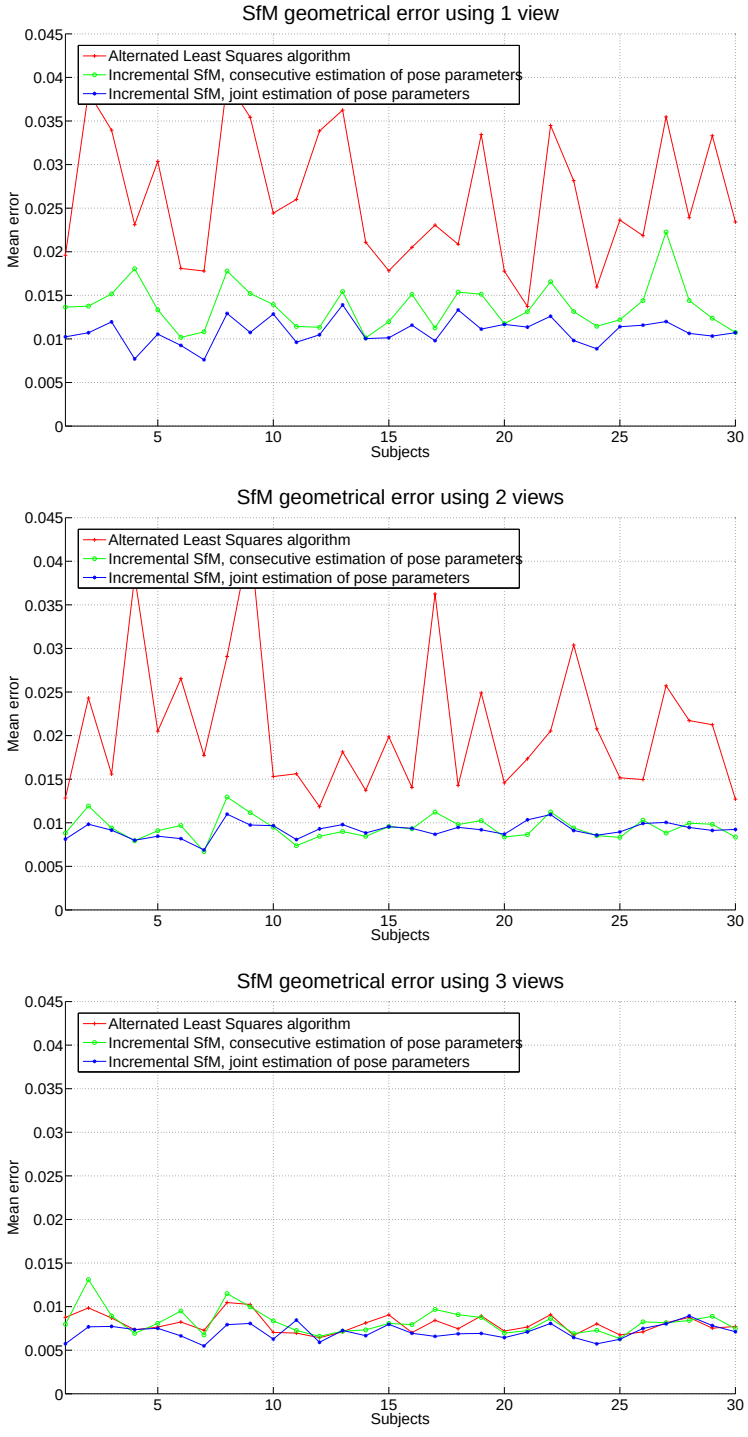

Figure 9: Geometrical error of the recovered 3D shapes for different individuals (**smile expression**), using 1 view (*top*), 2 views (*middle*) and 3 views (*bottom*)

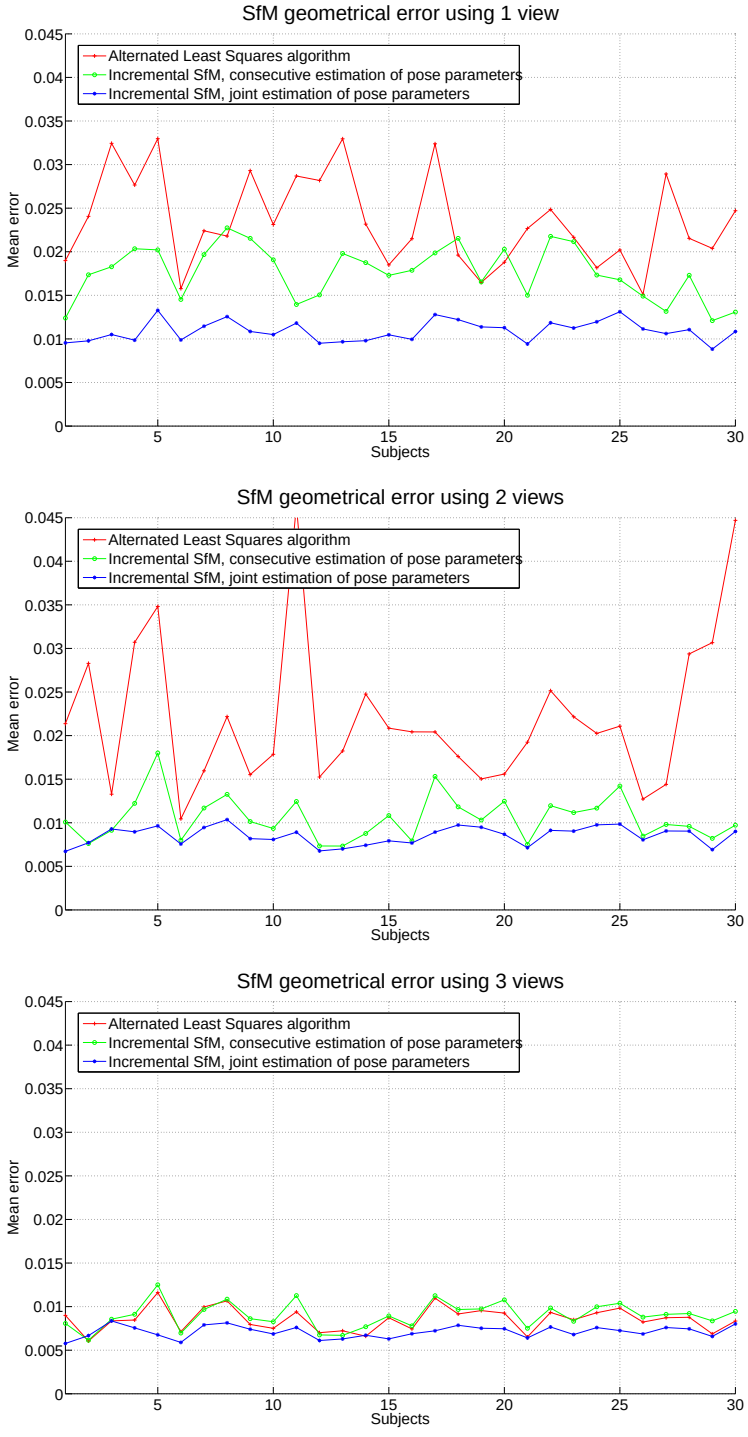

Figure 10: Geometrical error of the recovered 3D shapes for different individuals (**scream expression**), using 1 view (*top*), 2 views (*middle*) and 3 views (*bottom*)

The initialization point for the different SfM methods (and the prior for our algorithm) is obtained from a reduced set of 3D faces with neutral expression previously introduced. We apply PCA to these 16 preexisting face representations and select the 6 principal shape deformation vectors to generate a linear basis that is used as the initial value for the shape estimation and as regularizing term in our algorithm.

As it is shown in figures 8, 9 and 10, our incremental SfM algorithm obtains best average results in terms of reconstruction error. Even when only one view per person is provided, the algorithm is able to incorporate depth information from the prior dataset to reconstruct realistic 3D models. Having one view per subject may be interpreted as an extreme case for conventional SfM algorithms. However, in different applications (i.e. bioidentification based on identity cards) it is a common scenario for which our proposed technique must provide accurate solutions. As expected, when adding extra views, the algorithm reduces the geometrical error by consistently integrating this additional information. When the number of views is increased the gain provided by our algorithm gets smaller since the geometrical information provided by the training 2D database becomes more important compared to the prior.

Also it worth pointing out that the method using joint pose parameter estimation is more effective than the consecutive one. As mentioned in section 3.3, the joint method estimates a basis containing several "motion" vectors that linearly approximates changes in the shape projections due to pose changes in an interval around current pose estimate (tangent space). The estimated shape deformation basis will be orthogonal to this motion basis and the resulting linear deformation model is less biased by pose changes components present in training data.

When using only one training view, the use of an ALS algorithm without prior information causes a great variance in the geometrical error. The reason behind this behavior is that for some individuals the ALS algorithm recovers a degenerate scene configuration (analogous to the one shown in figure 2 (a)). Additionally, it is interesting to note that the errors achieved for the "neutral" expression case are smaller than "smile" or "scream". This is coherent with the dataset, as the shapes used for initialization (and prior information) also represent neutral expression faces and we can assume the starting point in the optimization is closest to the real desired solution.

The position of the landmarks in the different training images is determined by a manual labeling process. As a result, there is an error in the point locations that will affect the SfM performance. One consequence of this noisy training set (together with the simplifications in the camera model) is the residual error component at the convergence point that is presented in figures 8, 9 and 10. The more views we use in the model construction the more accurate the least squares estimate of the 3D shape is. This estimate minimizes the reprojection error, but it will never be zero due to the labeling errors (and the approximated camera model).

In the performed experiments, the reconstruction error has no significative variance for different landmark locations. As an example, figure 11 plots the

mean reconstruction error across the different subjects for each one of the 68 landmarks. The shape instances represented in this figure were obtained using our incremental SfM algorithm and 3 training views (0^o, 45^o and -60^o). As expected, the highest error values correspond to the "scream" expression and specially those points around the jaw contour (since these points show the biggest dissimilarity with respect to the closed mouth of the neutral expression used as initialization value).

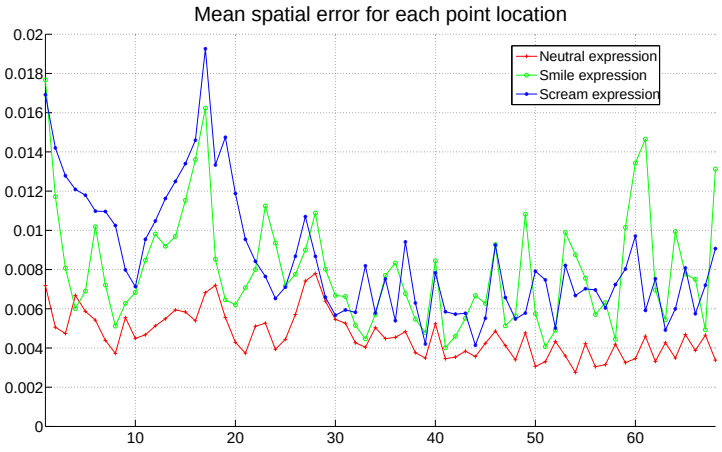

Figure 11: Mean geometrical error at each point location for "neutral", "smile" and "screen" expressions using our incremental SfM (3 views)

Figure 12 shows some examples of the obtained 3D models for the experiment using 3 training views after the densification and texture extraction process described in section 4. The dense point information was extracted from the 16 neutral subjects obtained from [13].

5.2. Generic face models using rank constraints

Similarly to other approaches, the proposed non-rigid SfM algorithm is based on the assumption that all possible shape configurations for the considered object are caused by a limited range of deformations that can be modeled using a reduced-rank basis. As we have seen in previous section, this restriction allows to learn an accurate shape description and pose estimates for training subjects using a set of noisy 2D projections. But it is also desirable that the learned shape basis could characterize the intrinsic non-rigid object structure of the object class so it can effectively generalize across new shape instances. In this section, we empirically show how our incremental SfM algorithm is able to model human faces having a defined expression, and generalize to previously unseen subjects.

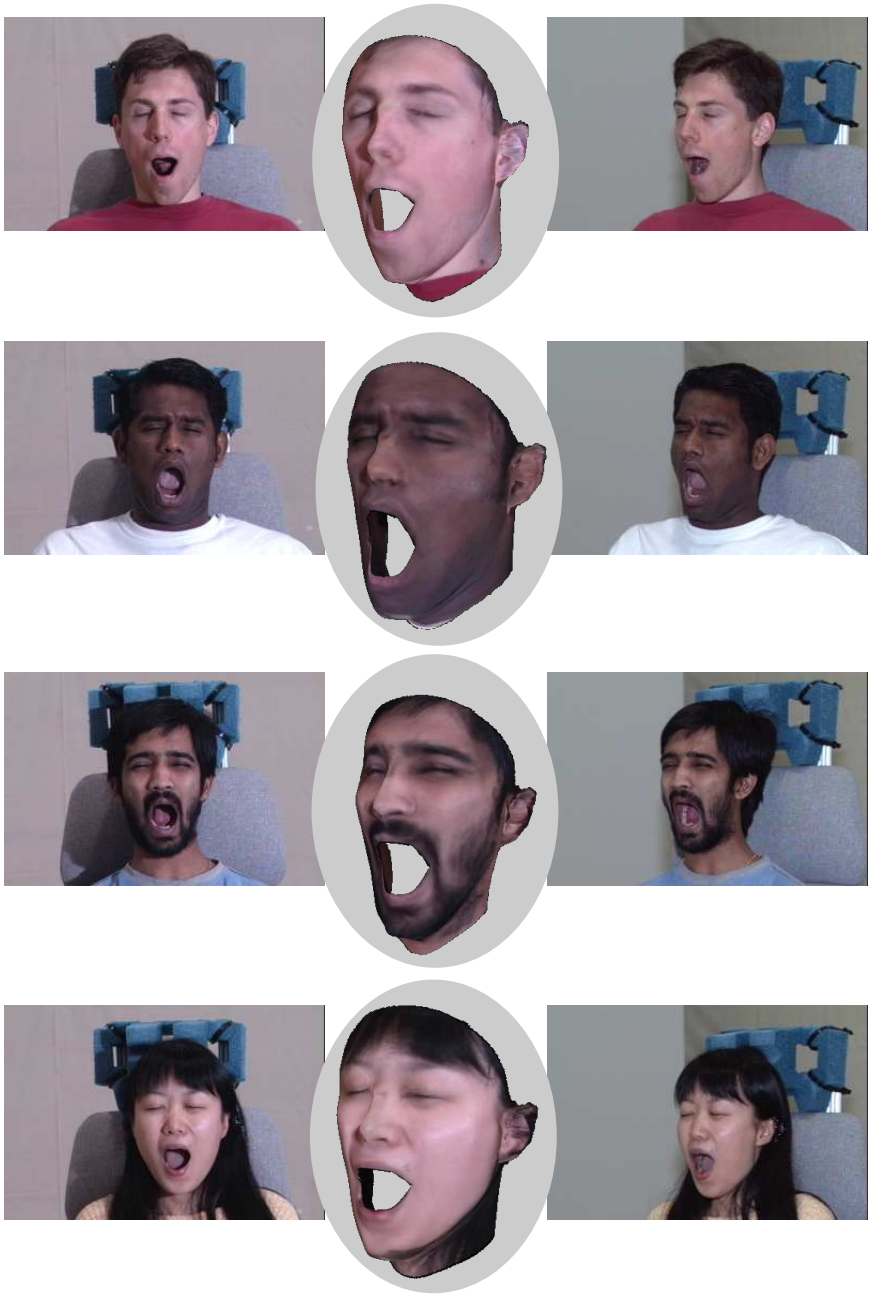

Figure 12: Mean geometrical error at each point location for “neutral”, “smile” and “screen” expressions using our incremental SfM (3 views)

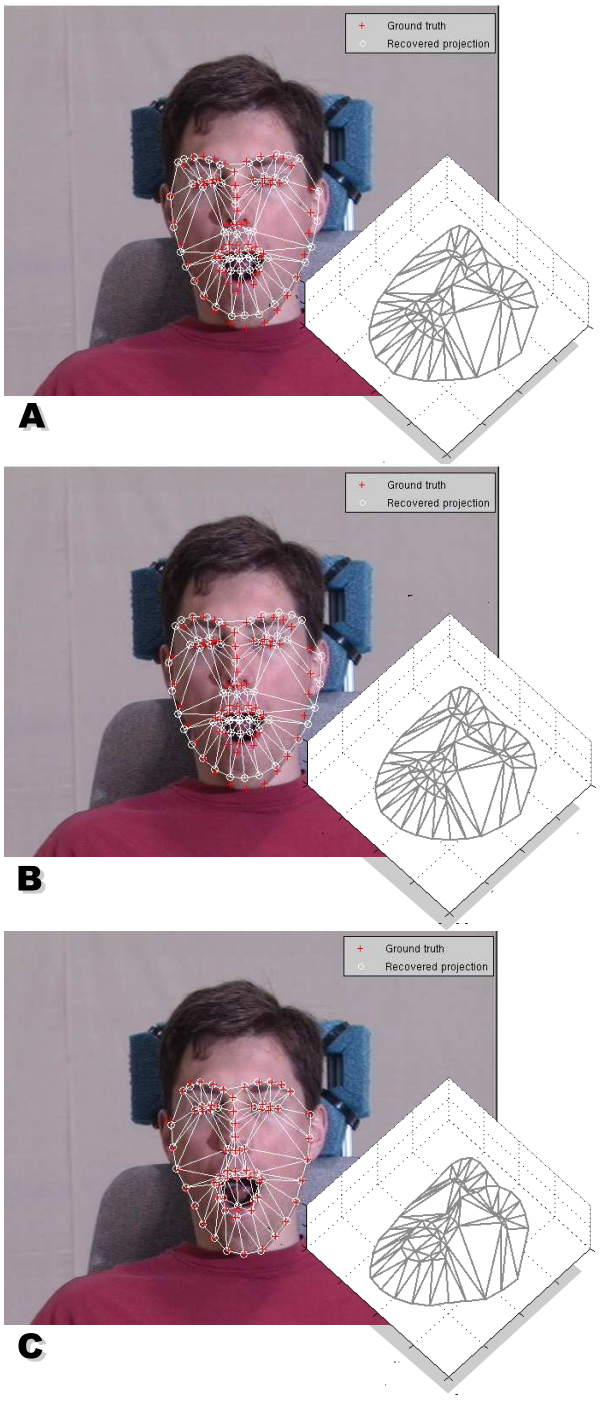

Figure 13: A: Fitting a morphable model to a novel view not included in the linear span of the 3D basis. B: Fitting using a regularized model. C: Fitting model obtained using incremental SfM algorithm.

In order to analyze the capability of the 3D deformable shape model to represent a new individual we need to determine which are the best parameters for the shape basis that generate a 3D representation of the new considered subject as accurate as possible. These optimal parameters can be obtained using a model fitting algorithm. In this section we use a gradient descent fitting algorithm described in [42] that minimizes the reprojection error and allows for regularization term. For each testing subject we provide a frontal view and the fitting algorithm solves for the optimal pose (motion) and shape parameters that describe that face assuming a RTS camera model.

Given a subject with an untrained expression, it is likely that the motion and shape parameters are biased in the fitting process. This might lead to reconstructed shapes far from the average and heavily distorted motion parameters. Figure 13 illustrates this situation. In figure 13 *(a)* we use the gradient descent algorithm to estimate the rigid pose parameters and shape coefficients of a face with a "surprise" expression using a 3D model learned exclusively from neutral expression instances. As it can be observed, the neutral model is not able to accurately reproduce a 3D mesh with open mouth. Figure 13 *(b)* represents the result of fitting the same linear 3D shape basis using a regularized approach, forcing shape parameters close to zero[1]. The estimated frontal projection is slightly less accurate than figure 13 *(a)* because proximity to the mean shape is imposed. However, the recovered 3D mesh presents less shape artifacts when projected onto a novel view. In this paper, we have proposed an incremental SfM algorithm that can be used to incorporate extra modes in the shape basis to represent a new expression. Figure 13 *(c)* represents the result achieved when fitting a model that has been learned using our SfM approach incorporating information about the "scream" expression from a 2D database.

In the following we provide a quantitative evaluation of the accuracy of the learned model to represent an "expression class" and generalize to previous unseen subjects. We generate different models representing "neutral", "smile" or "scream" expressions using our incremental SfM algorithm. These models are then fitted to a frontal view of a new subject and we measure the geometrical error between the generated shape instance and the 3D ground truth representation of that subject. The training dataset used to build these models is composed of 40 subjects and the same prior information as previous section (consisting in 16 3D shapes of faces having "neutral" expression). Different number of views are used in the experiments to compare the results (using the same rotation angles 0^o, 45^o and -60^o). The testing set consists of 30 different individuals having the desired expression from the CMU Multi-PIE database.

[1] The well known bias-variance analysis [49] shows that a trade-off between matching quality and prior density is needed in order to achieve good generalization performance. Based on this idea, Blanz et al. [42] proposed a regularized fitting process that prevents overfitting using a Bayesian formulation. Regularized approaches bias the estimate of the recovered shape towards the mean shape producing more realistic results. However, these approaches are unlikely to recover previously untrained facial expressions not contained in the linear span of the model.

The mean geometrical error between the reconstructed shape and ground truth is measured in order to evaluate the accuracy of the model when representing new shapes. For comparison purposes, we plot the error for the 3D shapes obtained by fitting the prior model (created performing PCA to the initial 16 neutral expression shapes) using the basic gradient descent fitting algorithm and its regularized version. Results are represented in figure 14. The reduced prior model is able to represent the subjects with neutral expression with relatively low error (top image). However, it fails to reproduce scream expressions at the bottom because it does not include deformation modes modeling the facial configuration for this expression (i.e. open mouth).

6. Conclusions

Most common approaches to build 3D Morphable Models use information extracted from laser scans. The amount of captures and the expression variability contained in those scans is usually limited and, in many cases, restricted to neutral expressions. We have illustrated how models trained exclusively from neutral expression databases do not generalize properly to new expressions involving significative changes in facial morphology (i.e. "scream" expression). On the other hand, we have proposed an incremental SfM algorithm to build generic 3D Morphable Models that is able to incorporate new deformation modes corresponding to different expressions from widely used images databases.

The main contributions of the paper are: (1) a method based on SfM to learn generic 3D face models from existing 2D databases, (2) the introduction of an incremental approach to incorporate prior 3D shape information in the SfM formulation. The proposed technique requires only a reduced number of views to accurately build the 3D models. No smoothness assumptions between different frames are needed (i.e. temporal continuity in the landmark sets) and, consequently, the algorithm can be used in wide baseline scenarios. The use of prior information in the form of preexisting 3D surfaces makes it possible to apply the algorithm to quasi-frontal image datasets with limited depth information and prevents degenerate solutions. The ALS based formulation deals with noise in manually labeled databases and missing data due to self-occlusions.

Experimental results show how our algorithm produces accurate shape reconstructions using one or a very reduced number of input images per subject. Comparative results show how our algorithm outperforms existing approaches for SfM in these scenarios with a limited number of training views. Finally, we have empirically verified that the rank constrained optimization involved in our algorithm yields generic models able to represent new shape instances for individuals not included in the training set.

Acknowledgments

This work was partially supported by the Ministry of Education and Science (CICYT) of Spain under contract TIN2006-01078 and Junta de Andalucia under

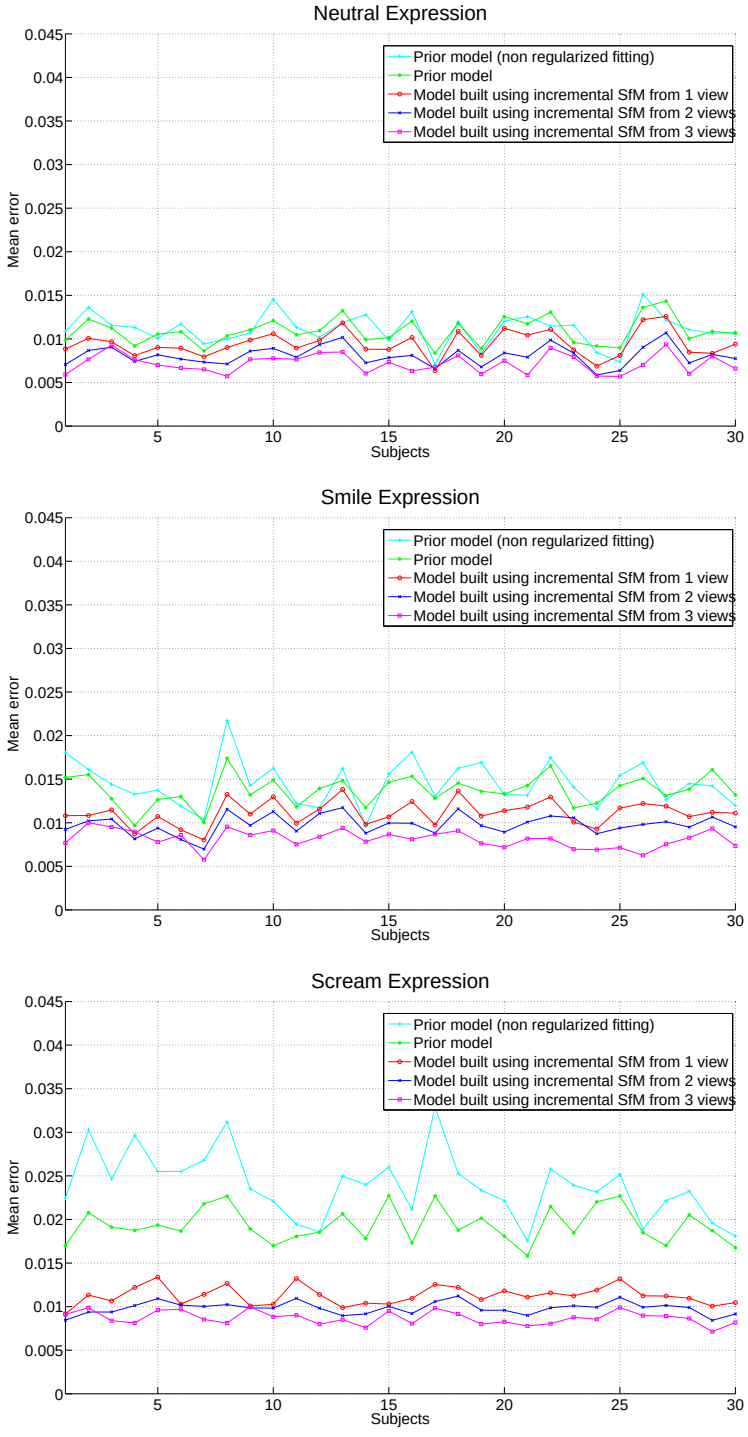

Figure 14: Mean geometrical errors using different linear shape models to represent 30 previously unseen 3D faces. Top: neutral expression. Middle: Smile expression. Bottom: Scream expression

contract TIC-02800. Thanks to the University of South Florida (USF) and the University of Freiburg for providing the Human ID 3D Database and the 3D Morphable Model. Thanks to Jeff Cohn for providing a partial labeling of the Multi-PIE database. Thanks to L. Torresani for providing a publicly available implementation of their SfM algorithm.

References

[1] V. Blanz, C. Basso, T. Poggio, T. Vetter, Reanimating faces in images and video, in: Computer Graphics Forum, Vol. 22, 2003, pp. 641–650.

[2] G. Edwards, T. Cootes, C. Taylor, Face recognition using active appearance models, in: ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume II, Springer-Verlag, London, UK, 1998, pp. 581–595.

[3] V. Blanz, T. Vetter, Face recognition based on fitting a 3d morphable model, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 9, 2003, pp. 1063–1074.

[4] S. Ramanathan, A. Kassim, Y. Venkatesh, W. Wah, Human facial expression recognition using a 3d morphable model, in: ICIP06, 2006, pp. 661–664.

[5] S. Baker, I. Matthews, J. Schneider, Automatic construction of active appearance models as an image coding problem, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (10) (2004) 1380–1384.

[6] M. J. Black, A. D. Jepson, Eigentracking: Robust matching and tracking of objects using view-based representation, International Journal of Computer Vision 26 (1) (1998) 63–84.

[7] I. Matthews, S. Baker, Active appearance models revisited, International Journal of Computer Vision (60) (2004) 135–164.

[8] S. Baker, I. Matthews, J. Schneider, Automatic construction of active appearance models as an image coding problem, IEEE transactions on Pattern Analysis and Machine Intelligence (26) (2004) 1380–1384.

[9] T. F. Cootes, G. V. Wheeler, K. N. Walker, C. J. Taylor, Coupled-view active appearance models, in: British Machine Vision Conference, 2000, pp. 52–61.

[10] F. de la Torre, J. Vitrià, P. Radeva, J. Melenchón, Eigenfiltering for flexible eigentracking, in: International Conference on Pattern Recognition, 2000, pp. 1118–1121.

[11] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.

[12] F. De la Torre, M. H. Nguyen, Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[13] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: Siggraph, 1999, pp. 187–194.

[14] L. Gu, T. Kanade, 3D alignment of face in a single image, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2006, pp. 1305–1312.

[15] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: Siggraph, 1999, pp. 187–194.

[16] C. Bregler, A. Hertzmann, H. Biermann, Recovering non-rigid 3D shape from image streams, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2000.

[17] M. Brand, Morphable 3D models from video, in: International Conference on Computer Vision and Pattern Recognition, 2001, p. 456463.

[18] J. Xiao, J. Chai, T. Kanade, A closed-form solution to non-rigid shape and motion recovery, International Journal of Computer Vision (67) (2006) 233–246.

[19] M. Brand, A direct method for 3D factorization of nonrigid motion observed in 2D, in: Internation Conference on Computer Vision and Pattern Recognition, 2005.

[20] L. Torresani, A. Hertzmann, C. Bregler, Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (5) (2008) 878–892.

[21] L. Torresani, A. Hertzmann, C. Bregler, Learning non-rigid 3D shape from 2D motion, in: Neural Information Processing Systems, 2003.

[22] P. Torr, A.W.Fitzgibbon, A. Zisserman, The problem of degeneracy in structure and motion recovey from uncalibrated image sequences, International Journal of Computer Vision 32 (1999) 27–44.

[23] A. Buchanan, A. Fitzgibbon, Damped newton algorithms for matrix factorization with missing data, in: International Conference on Computer Vision and Pattern Recognition, Vol. 2, 2005, pp. 312–322.

[24] D. Nister, F. Kahl, H. Stewenius, Structure from motion with missing data is NP-hard, in: International Conference on Computer Vision, 2007, pp. 1–7.

[25] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, International Journal of Computer Vision (9) (1992) 137–154.

[26] M. Maruyama, S. Kurumi, Bidirectional optimization for reconstructing 3D shape from an image sequence with missing data, in: IEEE International Conference on Image Processing, 1999, pp. 120–124.

[27] R. Guerreiro, P. Aguiar, 3D structure from video streams with partially overlapping images, in: IEEE International Conference on Image Processing, 2002, pp. 897–900.

[28] L. Torresani, D. Yang, E. Alexander, C. Bregler, Tracking and modeling non-rigid objects with rank constraints, in: International Conference on Computer Vision and Pattern Recognition, 2001.

[29] C. Julia, A. Sappa, F. Lumbreras, J. Serrat, A. Lopez, Factorization with missing and noisy data, in: Internation Conference on Computational Science, 2006.

[30] A. Buchanan, Investigation into matrix factorization when elements are unknown (2004).

[31] A. D. Bue, F. Smeraldi, L. Agapito, Non-rigid structure from motion using nonparametric tracking and non-linear optimization, in: IEEE Workshop in Articulated and Nonrigid Motion held in conjunction with CVPR2004, 2004.

[32] S. Olsen, A. Bartoli, Using priors for improving generalization in non-rigid structure-from-motion, in: British Machine Vision Conference, 2007.

[33] A. D. Bue, A factorization approach to structure from motion with shape priors, in: International Conference on Computer Vision and Pattern Recognition, 2008.

[34] J. Xiao, B. Georgescu, X. Zhou, D. Comaniciu, T. Kanade, Simultaneous registration and modeling of deformable shapes, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 2429–2436.

[35] J. Xiao, S. Baker, I. Matthews, T. Kanade, Real-time combined 2D+3D active appearance models, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004, pp. 535–542.

[36] I. Matthews, J. Xiao, S. Baker, 2D vs. 3D deformable face models: Representational power, construction and real-time fitting, International Journal of Computer Vision 75 (1) (2007) 93–113.

[37] X. Lu, A. K. Jain, Deformation modeling for robust 3D face matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (8).

[38] F. Al-Osaimi, M. Bennamoun, A. Milan, An expression deformation approach to non-rigid 3D face recognition, International Journal of Computer Vision.

[39] Y. Kim, S. Chung, B. Kim, S. Cho, 3D face modeling based on 3D dense morphable face shape model, International Journal of Computer Science and Engineering 2 (3).

[40] F. de la Torre, M. J. Black, A framework for robust subspace learning, International Journal of Computer Vision. 54 (2003) 117–142.

[41] D. Jacobs, Linear fitting with missing data: Applications to structure-from-motion and to characterizing intensity images, in: International Conference on Computer Vision and Pattern Recognition, 1997.

[42] V. Blanz, A. Mehl, T. Vetter, H. Seidel, A statistical method for robust 3D surface reconstruction from sparse data, in: International Symposium on 3D Data Processing, Visualization, and Transmission, 2004, pp. 293–300.

[43] M. Brand, Incremental singular value decomposition of uncertain data with missing values, in: European Conference on Computer Vision, 2002.

[44] D. Skočaj, A. Leonardis, Incremental and robust learning of subspace representations, Image and Vision Computing 26 (2008) 27–38.

[45] A. Levy, M. Lindenbaum, Sequential karhunen-loeve basis extraction and its application to images, IEEE Transactions on Image Processing 9 (8) (2000) 1371–1374.

[46] F. Bookstein, Principal warps: Thin-plate splines and the decomposition of deformations, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (1989) 567–585.

[47] J. Foley, A. van Dam, Fundamentals of interactive computer graphics. The systems programming series, 1984.

[48] R. Gross, I. Matthews, J. Cohn, S. Baker, Guide to CMU Multi-Pie face database (2002).

[49] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1996.