3D Spectral Nonrigid Registration of Facial Expression Scans

Gabriel L. Cuendet, *Student member, IEEE,* Christophe Ecabert, Marina Zimmermann, *Student member, IEEE,* Hazım K. Ekenel, and Jean-Philippe Thiran, *Senior Member, IEEE*

Abstract—In this paper, we introduce a new template-based spectral nonrigid registration method in which the target is represented using *multilevel partition of unity* (MPU) implicit surfaces and the template is embedded in a discrete Laplace-Beltrami based spectral representation using the *manifold harmonics transform* (MHT). The implicit surface parametrization of the target allows us to avoid computing correspondences during the registration as in classical nonrigid iterative closest point (ICP) techniques. It also allows us to denoise the 3D scans and fill the holes by interpolating the noisy 3D data and to incorporate different types of 3D surfaces into our model, independently of their original parametrization. We take advantage of spectral geometry processing methods to compute a spectral embedding of the template and use it as a parametric surface deformation model. We optimize the nonrigid deformation directly in the spectral domain, thus effectively reducing the size of the parameter space as compared with the classical *per vertex affine transformation* deformation model. In addition, we introduce a new 3D facial expressions database, EPFL3DFace, on which we apply the proposed method to nonrigidly register 3D face scans that contain different expressions. This database consists of 3D scans of 120 subjects posing 35 different facial expressions. These include various standard prototypical facial expressions as well as individual action units, visemes, and the facial movement of biting one's own upper lip, which are suitable for a large variety of applications.

Index Terms—3D face analysis, Nonrigid registration, 3D facial expressions database, Mesh deformation, Computational geometry, Spectral mesh processing

1 Introduction

Facial image analysis and synthesis have attracted a significant amount of attention in the last two decades from the computer vision and computer graphics research communities. These two communities have both tackled different but related problems: face recognition [1] [2] [3], head pose estimation [4], gaze tracking [5] [6], visual speech recognition, [7] facial expression recognition [8] [9] [10] [11], synthesis of 3D faces [12], facial animation [13] [14] [15], and face or expression transfer [16] [17].

The approaches that address these problems can benefit from the availability of low-cost 3D scanners such as the Microsoft Kinect® and take advantage of 3D facial images and 3D face models to avoid limitations inherent to 2D images such as self occlusions or sensitivity to head pose variations. Building a complete 3D face model from the ground up is still very demanding as the amount of data required to obtain a model which takes into account a large amount of variations in terms of identity and facial expressions is high and not easily available from public databases. The variance in appearance is influenced by factors such as age, gender and ethnicity, and when also taking facial expression variations into account, sampling the space of combinations of all these variations simply becomes intractable.

A certain number of databases consisting of 3D representations of the face have been proposed. An important difference between the databases is whether or not the 3D shapes share a common parametrization. Tasks like synthesis of 3D faces or facial animation require a generative model of shapes. These generative models must be learned from a database of consistently parametrized, i.e. registered, instances. Thus, the main challenge in constructing a generative model is to re-parameterize the example surfaces such that semantically corresponding points, e.g. the nose tips or mouth corners, share the same location in the parametrization domain. Existing 3D face models where 3D scans are registered and statistical analysis is performed include the MPI 3D Morphable Model (3DMM) [18], the multilinear face model [19], the Basel Face Model, [20], FaceWarehouse [21], the Large Scale Facial Model (LSFM) [22], the Surrey Face Model (SFM) [23] and the Robust Multilinear Model (RMM) [24], but amongst these, only FaceWarehouse and the RMM are trained with a large number of subjects and different facial expressions. These 3D face models are learned from large databases of 3D facial surfaces, containing representative examples spanning the range of variations that the model will be able to capture. As an example, a model learned only from 3D surfaces of neutral faces will not fit well on expressive faces nor be able to capture the variation between a smiling face and a sad face.

In this work, we contribute to the availability of more 3D facial surfaces by introducing EPFL3DFace, a new database consisting of 120 subjects performing 35 expressions. We show that the subspace spanned by our 120 subjects, among which 87% are Caucasian, extends the subspace spanned by the subjects from FaceWarehouse [21], another publicly available database of fully registered 3D facial scans includ-

- *G. Cuendet, C. Ecabert, M. Zimmermann and J.-Ph. Thiran are with the Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.*
  E-mail: firstname.lastname@epfl.ch
- *H. Ekenel is with Istanbul Technical University, Turkey.*

ing a variety of facial expressions.

In order to allow for statistical modeling, for example with a morphable model, a multilinear model, a blendshape model, etc., the scanned 3D facial surfaces have to be put into dense correspondence by nonrigidly registering the 3D surfaces. The general strategy is for each scan to deform a template, the *floating surface*, or *source* $\mathcal{S}$ such that it matches the scan or *target surface* $\mathcal{T}$. The template parametrization is thus transferred to each of the scans. In this work, we propose to compute a spectral embedding of the source and use that representation to constrain the possible deformations. Deforming the source in the spectral domain allows to choose which frequency band to focus on, depending on required properties. In our case, the deformation model obtained by embedding the source in the spectral domain allows us to optimize the deformation over the parameters corresponding to low frequencies and enforce this deformation to be smooth. Moreover, this also presents the advantage of providing a compact deformation model as compared to *as-rigid-as-possible* deformation. Thus, the number of parameters to optimize can be kept small, as demonstrated in this work, where our deformation model is parameterized with approximately 100 times less parameters than the well known *per-vertex affine transform*. We also propose to represent the target as an implicit surface in order to avoid computing correspondences, when evaluating the distance between the source and the target and the gradient of that distance. This representation allows to approximate rather than interpolate the surface, which is beneficial in the case of noisy surfaces.

Establishing correspondences from one surface to another has been investigated in several fields and under different names such as *nonrigid registration*, *alignment*, *matching*, *mesh morphing*, *cross-parameterization* or *correspondence estimation*. A few of the most relevant methods are discussed hereafter and we refer the reader to the book of Bronstein et al. [25] or the surveys of Van Kaick et al. [26] and Tam et al. [27] for more exhaustive reviews of the different methods.

In the remaining of this paper, we first review some important related work in the fields of computer graphics and computer vision and provide a comparison with our work. Section 3 describes the new nonrigid registration method that we propose. We then introduce EPFL3DFace, our database of 3D facial expressions in section 4 and present results achieved by the proposed method on the new database in section 5. Finally section 6 summarizes the contributions of this work and discusses a few directions for future work.

2 Previous work

Blanz and Vetter first introduced the term *morphable model* [18] to describe their parametric face modeling technique based on a large number of 3D face scans. In order to establish correspondence between all individual face scans, they use cylindrical coordinates both for color and geometry information and adapted the optical flow algorithm to compute a vector field of displacement between points [33]. Their method is well suited for data acquired with a 3D scanner using cylindrical coordinates or that can easily be converted to that particular planar representation.

In [34], the authors present a template-based nonrigid registration method to compute dense point-to-point correspondence between surfaces with the same overall structure, but substantial variation in shape, such as human bodies. They formulate this as an optimization problem over a set of *per vertex* affine transformations. The objective function includes three terms: a data term defined as the sum of squared distances between spatially close vertices on the source and the target surfaces, a smoothness term which enforces that neighboring affine transformations are as similar as possible and a marker term defined as the sum of squared distances between a set of marker's locations on the template surface and on the target surface. By ensuring the smoothness of the transformations over the surface, they define an *as-rigid-as-possible per vertex affine transform* further constrained with a set of 3D marker locations. By using domain knowledge inherent in the template surface, this method is robust to incomplete surface data and is able to fill in holes or poorly captured parts of the surface.

Vlasic et al. [19] applied this template-fitting procedure to 3D face scans and described multilinear face models for expression transfer. In [35] Mpiperis et al. follow a method similar to [34] but add an error term looking for correspondences directed from the target surface to the source and not only in the other direction. They claim that this is important at the beginning of the optimization process when the source is far from the target and it helps avoiding local minima by making the resulting vector field smoother.

Extending the idea of iterative closest point (ICP) [36] to nonrigid registration and in particular defining optimal steps using a series of stiffness weights to regularize the deformation described in [34], Amberg et al. defined the optimal step Nonrigid ICP (NICP) [37]. They express the cost function as a least squares problem, thus being able to determine in each step of the algorithm the optimal deformation, in the sense that it exactly minimizes the cost function for fixed stiffness and correspondences.

Further extending the method, Cheng et al. proposed to incorporate a statistical shape prior [38] into the fitting procedure of NICP in order to avoid noisy fitting results and even non-face like fitting due to its weak constraint on the shape geometry. The statistical shape prior is a deformable 3D face model [39], [40], whose optimal controlling parameters are solved in an alternating manner. Along the same line, Brunton et al. [41] proposed a detailed review of statistical shape models. They emphasize that to incorporate a statistical shape model to fit to data, instead of a template-based approach with a nonrigid ICP approach and regularization constraints, can significantly reduce the search space. This results in the ability to reconstruct the underlying shape in the presence of severe noise or occlusions.

Weise et al. [13] also followed a nonrigid ICP approach, optimizing a cost function composed of three terms. Nevertheless, they introduced a combination of point-to-point distance and point-to-plane distance as discussed in [42] in the data-term and expressed the smoothness term as a membrane energy on the displacement vectors, using the standard cotangent discretization of the Laplace-Beltrami operator.

Sumner et al. [32] introduced an embedded deformation model composed of a collection of affine transformations

Name	subj.	expr.	v.	Acquisition/Source	landm.
3D Morphable Model (3DMM), MPI Tübingen [18]	200	Neutral	≈70k	Cyberware	
Spacetime Faces [28]	1	384	23.728k	Custom structured light scanner	
Multilinear face model [19]	15 + 16	10 + 10	≈30k	3dMD/3Q's	21
Human Face [29]	1	15	≈2k	Custom structured light scanner	
Basel Face Model [20]	200	Neutral	53.49k	ABW-3D	
FaceWarehouse [21]	150	20 (47)	11.51k	Microsoft Kinect®	74
Large Scale Facial Model (LSFM) [22]	9663	Neutral	53.215k	3dMD	
Surrey Face Model (SFM) [23]	169	Neutral	29.587k [1]	3dMDface	46
Robust Multilinear Model (RMM) [24]	205	23	5.996k	Bosphorus [30] & BU-3DFE [31]	
EPFL3DFace	120	35	11.51k	Microsoft Kinect ®	74

TABLE 1
Comparison of registered 3D face databases and 3D face models in terms of number of subjects (subj.), number of expressions (expr.), number of vertices of the aligned surfaces (v.), sensor used for data acquisition, and number of landmarks (landm.). Note that in the RMM, no new 3D data are recorded, but 3D data from the Bosphorus [30] and BU-3DFE [31] databases are registered using [32].

organized in a graph structure. One transformation is associated with each node of a graph embedded in $\mathbb{R}^3$, so that the graph provides spatial organization to the deformations. Each affine transformation induces a localized deformation on the nearby space. That approach was later adapted by Li et al. [43] to handle motion in the data. This nonrigid registration approach is successfully used for real-time performance-based facial animation in [15].

In [44] Zell et al. extended the nonrigid ICP approach to surfaces which cannot be considered near-isometric and for which the closest point correspondences might be invalid by first mapping the source and target surfaces into a simpler space and computing correspondences there. The simpler space is a smoothed, feature-less version of the input models computed by a joint fairing technique based on Laplacian smoothing. To compute correspondences, they iteratively minimize a cost function, which includes three terms: a data term and a marker term, similarly to previously described approaches, and a smoothness term defined as the norm of the Laplacians of vertex displacements, similar to the one used in [13].

Recently, Huber et al. released the *Surrey Face Model* (SFM) [23], a multi-resolution 3D morphable face model trained with 169 subjects with a neutral facial expression. Their nonrigid registration method was previously described in [45] and is an iterative coarse to fine method based on [46]. This method comprises three steps: first landmarks on the source and the target surfaces are brought into correspondence using thin plate spline (TPS) interpolation technique. Then, corresponding points on the source and the target are computed. The search for corresponding closest points takes into account not only the distance between points on the source and the target surfaces but also the angle between their normals, and the difference between curvature shape indices. Finally the positions of the source points are optimized in an *as-rigid-as-possible* fashion.

Bolkart et al. [24] emphasize the chicken-and-egg nature of the problem of training a new statistical face model: given a set of shapes and dense correspondences, a statistical model can be learned and given a representative model, better correspondences can be computed among a set of shapes. They propose a fully automatic approach to optimize the correspondences for 3D face databases based on multilinear statistical models using groupwise multilinear correspondences [24]. This method measures the model quality and optimizes the registration in such a way that the quality of both the model and the registration improve but an initial registration remains necessary. In their work, they first use a blendshape model to address the expression fitting problem. The 3D blendshapes were manually generated using a commercial software. To further nonrigidly deform the template corresponding to the correct expression, they use an embedded deformation framework [32]. This method was applied to two existing databases of 3D facial surfaces, the Bosphorus database [30] and the BU-3DFE database [31] and resulted in the Robust Multilinear Model (RMM) [24].

As an alternative to nonrigid ICP, some methods compute correspondences between two surfaces by embedding the intrinsic geometry of one surface into the other using generalized multi-dimensional scaling (GMDS) [47]. The good performance of this kind of methods has been demonstrated for face recognition and are an alternative to deal with variations due to facial expressions [48] [29]. As GMDS methods do not impose that close-by points on one surface map to close-by points on the other, the results are often spatially inconsistent.

In existing 3D facial expression databases, only FaceWarehouse, a 3D facial expression database for visual computing, released by Cao et al. [21], has both a large number of subjects and a variety of facial expressions. It consists of registered 3D surfaces of the head of 150 subjects performing 19 facial expressions plus a neutral face. The facial surfaces of the subjects were acquired with a Microsoft Kinect®. To register the 3D scans together, they used a two-step process, close to the nonrigid ICP methods described above. In the first step, Blanz and Vetter's morphable model [20] is automatically fitted and used as a parametric template. The nonrigid alignment between the fitted model and each of the neutral scans is then refined by allowing the obtained mesh to deform using a Laplacian-based mesh deformation algorithm [49]. Finally, the scans containing facial expressions are aligned using a deformation transfer algorithm [50] and refined with the same Laplacian-based mesh deformation algorithm.

Table 1 provides a comparison of EPFL3DFace, our new face expressions database with respect to existing 3D face models and databases in which the facial surfaces have been registered and are in dense correspondence with each other. In existing 3D facial expression databases, only FaceWarehouse [21] has both a large number of subjects

1. Multiresolution model with different levels of detail and number of vertices: 29.587k / 16.759k / 3.448k

and a variety of facial expressions. In comparison to that database, EPFL3DFace provides additional visemes suitable for visual speech recognition applications, additional facial expressions, and an extreme facial movement. In total, EPFL3DFace contains 35 scans for each subject, whereas FaceWarehouse contains 20 scans. In addition, FaceWarehouse and EPFL3DFace contain subjects from different populations, mostly Asian in FaceWarehouse and mostly Caucasian in EPFL3DFace, and can be considered as complementary in that respect.

In our nonrigid registration pipeline, we propose to embed the template in the spectral domain using a *manifold harmonics transform* (MHT) [51] and use this embedding as a surface deformation model. Indeed, by optimizing over the parameters corresponding to lower frequencies, we enforce the deformation to be smooth. Moreover, depending on the number of frequencies M_{freq} chosen, the number of parameters to optimize, $3 \times M_{\text{freq}}$, is much smaller than in the case of *per-vertex affine transform*, $12 \times N_{\text{vert}}$ as $M_{\text{freq}} < N_{\text{vert}}$. As an example, in our experiments, the template has $N_{\text{vert}} = 11510$ vertices. That would results in 138'120 parameters to optimize in a *per-vertex affine transform* model but our spectral embedding uses 500 basis functions, resulting in 1500 parameters to optimize in our transformation model, thus reducing the number of parameters by a factor 92.

A second keypoint of our method is the implicit surface representation [52] of the target 3D scans in order to overcome the problem of point correspondence. By representing the target as an analytical implicit surface, defined as the zero level-set of a squared distance function, the distance of any point to the surface is obtained by evaluating the value of the implicit function at that point. Moreover, when computing the implicit surface representation, the implicit function can approximate the original scan, thus effectively removing noise and filling holes.

3 METHODS

The complete alignment pipeline is composed of the following steps, described in detail in the following subsections: first, the template is rigidly aligned to the target such that both surfaces share the same scale, position and orientation in space. This initial rigid alignment is described in subsection 3.1. The different parts of the nonrigid registration are then described in subsection 3.2: the similarity measure using implicit surface representation in subsection 3.2.1, the transformation model using MHT in subsection 3.2.2 and the complete objective function and optimization process in subsection 3.2.3.

3.1 Initial rigid 3D scan registration

Our scans are, in general, not rigidly aligned with the template. Before being able to nonrigidly align the source to the target, it is essential to compensate for unknown rigid transformations such as scale, translation and rotation.

3D feature points, or landmarks, are used to compute the rigid transform between the source and the target such that the source is rigidly aligned to the target. First, 68 landmarks are manually annotated on the source. Note that this is done only once as the sources used for each expression are already registered.

Then, similar to the approach used in the LSFM [22], we automatically detect the same 68 landmarks on each target. An image is first generated by projecting the 3D surface on the image plane of a frontal virtual camera. We then detect the landmarks on this image using a state-of-the-art facial feature detection algorithm [53] based on the supervised descent method (SDM) [54]. In order to get the 3D positions of the landmarks on the target, we back-project the 2D positions of the landmarks with the known projection matrix of the virtual camera and intersect these rays with the 3D surface. The landmarks on the jaw are often less precisely located on the 3D surface due to the fact that the back-projected rays are almost tangential to the surface and thus a small imprecision in 2D becomes a large error in the intersection. For that reason, we discard these when computing the rigid transform.

Finally, the rigid transform, *i.e.* the translation and rotation between the two sets of 3D landmarks is computed as a weighted least-squares problem using a singular value decomposition (SVD) [55], [56]. The scaling factor between the two sets is retrieved as well. The scaling, translation, and rotation are applied to the source and the resulting shape is used for the nonrigid registration described in the next section.

3.2 Nonrigid 3D scan registration

Each scanned 3D facial surface needs to be re-parametrized into a consistent form, where the number of vertices, the triangulation, and the anatomical meaning of each vertex are consistent across all surfaces. The general strategy is for each scan to deform a rigidly aligned template, the *source*, $\mathcal{S}$ such that it matches the scan or *target surface*, $\mathcal{T}$. The deformation model, which ensures a meaningful deformation, is denoted by χ and the quality of the match is measured by a similarity measure.

$$\mathcal{S} = \{\boldsymbol{p}_i | i = 1, ..., N^S\} \overset{\chi}{\mapsto} \mathcal{T} = \{\boldsymbol{q}_i | i = 1, ..., N^T\}. \quad (1)$$

This dense correspondence problem is referred to as nonrigid registration and is defined by three main elements:

- a similarity measure, dependent on the representations of the *source* $\mathcal{S}$ and the *target* $\mathcal{T}$,
- a transformation model χ, which describes allowed deformations of the source, and
- an objective function, which combines the similarity measure and the transformation model and is optimized with a numerical optimizer.

In the next subsections 3.2.1 to 3.2.3, we will detail each of these three elements.

3.2.1 Similarity measure with implicit surface representation

In classical nonrigid ICP approaches, correspondences need to be computed in order to be able to evaluate the distance between the source and the target. These are unknowns as this is precisely what we are looking for in the first place. Several iterative approaches have been proposed based on spatial proximity of points, either using a *point-to-point* or a *point-to-plane* distance and looking for correspondences from

the source to the target or the opposite, or a combination of both [13] [42]. The correspondence problem gets even more complicated, when the quality of one or both surfaces is low. In particular, holes and noisy parts in the target further complicate the search for correspondence.

We propose to use an implicit surface representation for the target in order to avoid having to estimate correspondences. The surface is then implicitly represented as the zero level-set of a distance function $\mathfrak{d} : \mathbb{R}^3 \mapsto \mathbb{R}$. Choosing carefully that function allows to approximate the input surface rather than interpolate it, thus smoothing it and filling holes. In addition, desirable properties of an implicit surface reconstruction method include speed and low memory overhead.

As the value of the function is the signed distance to the surface, evaluating a distance between the source and the target can be achieved by simply summing the squared value of the implicit function at each vertex of the source, as described in equation (2). This does not require searching for correspondences.

$$\text{dist}^2(\mathcal{S}, \mathcal{T}) = \sum_i \mathfrak{d}\left(\boldsymbol{p}_i\right)^2. \tag{2}$$

Multilevel partition of unity (MPU) implicits provide fast, accurate, and adaptive reconstructions of complex shapes [52]. The main advantage of MPU is to define approximants locally, thus avoiding the overhead of a global support, and integrate them together by weighting each of them. Following the original method, we use the quadratic B-spline to generate weight functions.

MPU uses a hierarchical structure to adaptively divide the region of space containing the input set of shape vertices. We use an Octree structure, starting from the bounding cube of the shape and computing an approximation of the points enclosed in a sphere of radius R. The radius of the sphere is proportional to the main diagonal d of the current cell $R = \alpha d$. When the computed local max-norm approximation error ϵ is greater than a user-specified threshold ϵ_0, the cell is subdivided and the process is repeated. If the initial sphere does not contain enough points to compute the approximation, the radius is iteratively increased until the sphere contains a user-defined minimum number of points N_{min}. In that case, the cell is not further subdivided, independently of the approximation error and unlike the original method in which the initial sphere needs to be empty to stop the subdivision. The local max-norm approximation error ϵ is estimated according to the Taubin distance [57] and is given by equation (3).

$$\epsilon = \max_{|\boldsymbol{p}_i - \boldsymbol{c}| < R} |Q(\boldsymbol{p}_i)| / |\nabla Q(\boldsymbol{p}_i)|. \tag{3}$$

The choice of the approximants allows to address different scenarios: locally planar surfaces, surfaces with sharp edges, etc., as emphasized in [52]. Following the original method, we implemented the bivariate quadratic polynomial and the general quadric approximants. To give an intuition, the bivariate quadratic polynomial is best suited to approximate local smooth patches, and the general quadric provides consistent approximations on larger parts of the surface which might contain more than one sheet.

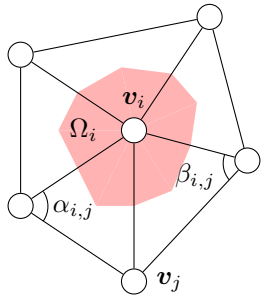


Fig. 1. Angles and local averaging area,Ω_i, used in the discrete Laplace-Beltrami operator

In practice, the surfaces we are implicitly representing, our scans, are mainly composed of local smooth patches in the region of interest, the face region, and noisy boundaries. Therefore, we only use the bivariate quadratic polynomial approximant. This and the choice of N_{min} have shown to be critical, when implicitly representing the scans from our database as explained in section 4.

3.2.2 Transformation model

When deforming the source toward the target, the transformation model defines the possible transformations of the source in order to avoid overfitting, prohibit arbitrary deformations, and favor reasonable ones and reduce the dimensionality of the problem. Intuitively, coarse, global deformations should be applied first and then refined with fine, local deformations. In general, smoothness should also be preserved.

Per-vertex displacements are thus modeled using spectral tools [58]. They offer an intuitive control over deformations where coarse, global deformations are embedded in the low frequencies and fine, localized deformations in the high frequencies. By selecting a number of lower frequencies $m \ll n$, the number of vertices in the source, the dimensions of the optimization problem are reduced. Moreover, the built-in smoothness of the low frequencies helps to avoid overfitting.

The Laplacian framework and differential representations allow to describe surface meshes through their differential properties. As a generalization of Fourier analysis the *Manifold Harmonics Basis* (MHB) and corresponding *Manifold Harmonics Transform* (MHT) introduced in [51] provide a reparametrization tool which allows us to represent a mesh with potentially fewer coefficients and more interestingly to constrain the deformation of the mesh, when changing the coefficients in ways that preserve the smoothness of the mesh.

Manifold harmonics are defined as the eigenfunctions of the discrete Laplace operator. The basis vectors of the MHT are thus the eigenvectors $\boldsymbol{h}^k$ of the discrete Laplacian as described in equation (4).

$$\boldsymbol{h}^k = [H_1^k, ..., H_n^k] \quad \text{satisfies} \quad -Q\boldsymbol{h}^k = \lambda D \boldsymbol{h}^k. \tag{4}$$

The matrix Q is called the *stiffness matrix* and is defined by the *cotangent formula*:

$$Q_{i,j} = \begin{cases} \frac{1}{2}\left(\cot(\alpha_{i,j}) + \cot(\beta_{i,j})\right) & \text{when } i \neq j \\ -\sum_k Q_{i,k} & \text{when } i = j. \end{cases}$$

where the angles $\alpha_{i,j}$ and $\beta_{i,j}$ are illustrated in figure 1.

The diagonal matrix D is called the *lumped mass matrix* and is defined by:

$$D_{i,i} = \sum_{t \in St(i)} \Omega_t,$$

where $St(i)$ denotes the set of triangles incident to i and Ω_t the local averaging area of triangle t. In our case, we use the barycentric cell as local averaging area. The barycentric cell connects the triangle barycenter with the edges' midpoints. The eigendecomposition of the discrete Laplacian described by equation (4) is computed using the band-by-band algorithm described in [51], which takes advantage of the *Shift-Invert* spectral transform.

To compute the transform of the function x from geometric space to frequency space, x is projected onto the manifold harmonics basis through the inner product. The MHT of x is a vector $[\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_m]$ given by equation (5).

$$\tilde{x}_k = < x, H^k > = \boldsymbol{x}^T D \boldsymbol{h}^k = \sum_{i=1}^{n} x_i D_{i,i} H_i^k. \quad (5)$$

The inner product contains D in order to ensure orthogonality of the basis, as the Laplacian is not symmetric, due to the weights $D_{i,i}$ which scale the lines of Q.

The inverse transform, to map the function $\tilde{x}$ in frequency space into its geometric space is given by equation (6).

$$x_i = \sum_{k=1}^{m} \tilde{x}_k H_i^k. \quad (6)$$

H is a basis containing the spectral modes of variation of the shape. We thus represent a new shape as the original source shape $\bar{\boldsymbol{p}}$ and a linear combination of spectral deformations, as described in equation (7).

$$\boldsymbol{p}(\boldsymbol{\alpha}) = \bar{\boldsymbol{p}} + H\boldsymbol{\alpha}, \quad (7)$$

where $\boldsymbol{\alpha}$ is a vector of spectral coefficients. Setting $\boldsymbol{\alpha}$ to zero yields the initial shape, without deformation.

Furthermore, as described in section 3.1, the source has been rigidly aligned to the target beforehand. Nevertheless, as pointed out by Blanz et al. [59], the result of this rigid prealignment is sub-optimal, since the optimal rigid alignment depends on the source after deformation. Thus we need to include translation and rotation in the transformation model. Translation is included in the first spectral basis, which is a constant vector, and we include a linearized rotation similarly to [59] as described in equation (8).

$$\begin{aligned} \boldsymbol{R}\boldsymbol{v} \approx & c_\gamma \boldsymbol{s}_\gamma + c_\theta \boldsymbol{s}_\theta + c_\phi \boldsymbol{s}_\phi + \boldsymbol{v} \quad (8)\\ \boldsymbol{s}_\gamma = & (-y_1, x_1, 0, -y_2, x_2, 0, ...)^T \\ \boldsymbol{s}_\theta = & (0, -z_1, y_1, 0, -z_2, y_2, ...)^T \\ \boldsymbol{s}_\phi = & (z_1, 0, -x_1, z_2, 0, -x_2, ...)^T \end{aligned}$$

The complete transformation model is thus given by equation (9).

$$\boldsymbol{p}(c_\gamma, c_\theta, c_\phi, \boldsymbol{\alpha}) = \bar{\boldsymbol{p}} + c_\gamma \boldsymbol{s}_\gamma + c_\theta \boldsymbol{s}_\theta + c_\phi \boldsymbol{s}_\phi + H\boldsymbol{\alpha}. \quad (9)$$

3.2.3 Objective function

Combining the transformation model and the implicit surface distance measure, we can evaluate the similarity between the deformed source and the target for a given set of parameters $\boldsymbol{\alpha}$, c_γ, c_θ, c_ϕ. We define the data fitting term E_{data} of our objective function as in equation (10).

$$E_{data} = \mathfrak{d}^t \left(\bar{\boldsymbol{p}} + c_\gamma \boldsymbol{s}_\gamma + c_\theta \boldsymbol{s}_\theta + c_\phi \boldsymbol{s}_\phi + H\boldsymbol{\alpha} \right). \quad (10)$$

We noticed that, due to the relatively low accuracy of the Kinect, the eye regions often do not contain enough details to correctly align the eyes. This causes the eyes of the source to slide on the flat region around the eyes of the target surface, ending in incorrect positions. To further constrain the eye regions, we use 3D landmarks around the eyes. On the target, these landmarks are detected with high accuracy during the rigid alignment step, whereas on the source, they have been manually annotated. The landmarks detection and annotation process is detailed in section 3.1. In order to constrain the eye regions, we add a term to the objective function penalizing large distances between the landmarks on the source and the corresponding landmarks on the target. This term is defined in equation (11).

$$E_l = \sum_{i=1}^{n_l} \|\hat{p}_i - \hat{q}_i\|_2^2, \quad (11)$$

where n_l is the number of landmarks, $\hat{p}_i$ are the landmarks on the source and $\hat{q}_i$ are the landmarks on the target.

As discussed in section 3.2.2, we want to favor low frequencies over high frequencies, thus we add a regularization term E_b to penalize higher bending of the deformation. This regularization term is defined in equation (12).

$$E_b = \|\Lambda_H \boldsymbol{\alpha}\|_2^2, \quad (12)$$

where Λ_H is a diagonal matrix of eigenvalues corresponding to the spectral bases.

A second regularization term E_m penalizes the magnitude of the deformation, as defined in equation (13).

$$E_m = \|\boldsymbol{\alpha}\|_2^2. \quad (13)$$

The complete objective function is given in equation (14).

$$E = E_{data} + \beta_0 E_l + \beta_1 E_b + \beta_2 E_m. \quad (14)$$

We use a gradient descent solver to minimize E. In our experiments, we chose $\beta_0 = 1e^{-4}$, $\beta_1 = 2e^{-3}$ and $\beta_2 = 2e^{-4}$ empirically.

4 EPFL3DFACE DATABASE

We have collected EPFL3DFace, a new 3D facial expressions database, for the study. The 3D facial surfaces have been nonrigidly registered with the method presented such that they are all in dense correspondence. This allows the use of EPFL3DFace database to train a 3D statistical model of the face, for example a morphable model, a multilinear model or a blendshape model, for a large variety of applications, such as but not limited to facial expression recognition, visual speech recognition, morphological analysis of the face, etc.

We recorded 120 subjects performing 35 facial expressions, while sitting still on a rotating chair. The subjects were

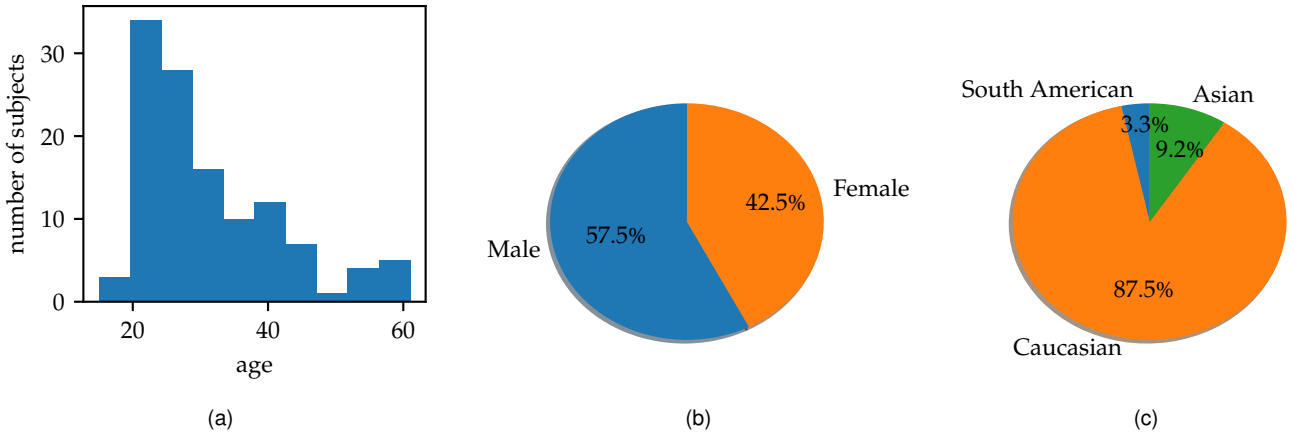


Fig. 2. (a) Age, (b) gender and (c) ethnicity distributions of the subjects included in the database

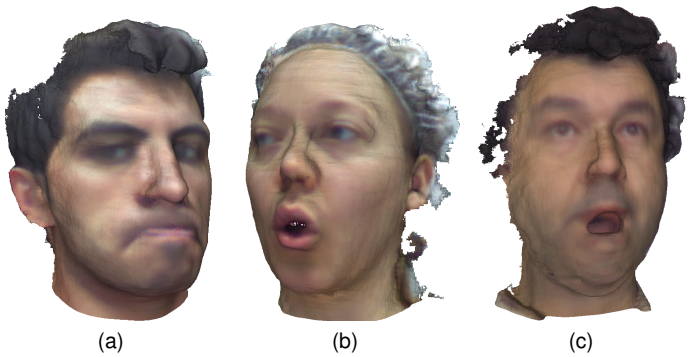


Fig. 3. Examples of scans from the database: (a) jaw forward (AU29), (b) viseme /uh/, (c) surprise

facing a Microsoft Kinect® for Windows v.1 at a distance of 50-70 cm. A screen in front of the subjects was displaying instructions on how to perform each expression with visual examples. At the same time, an operator was explaining and demonstrating how to perform the expression. Each subject had to perform each expression and stay perfectly still, while the operator was rotating the chair at an angle of $\pm 60^\circ - 90^\circ$. This operation took approximately 15 seconds on average.

Figure 2 shows the age, gender, and ethnicity distributions of the subjects included in EPFL3DFace database. In general, the population is slightly biased towards young men, as the subjects were recruited mainly in the electrical engineering department of the university. With 43% women and 57% men, the gender distribution is still reasonably well balanced. The ethnicity is strongly biased towards Caucasian, with 87% of the subjects included in the database. This is a wanted feature of the database making it complementary to FaceWarehouse [21], which mainly contains Asian subjects. We discuss this aspect in more detail in section 5.2. We also recorded the country of origin and the mother-tongue of the subjects.

We recorded each subject with a neutral facial expression, with the eyes open, and then instructed them to perform different facial expressions. These include prototypical expressions: anger, sadness, surprise, fear, disgust, happiness, and variants: anger with mouth slightly open, sad surprise and grin. They also include specific action units (AU): closed eyes (AU43), mouth open (AU25), brow lower (AU04), brow raiser (AU01), jaw left and right (AU30), jaw forward (AU29), mouth left and right, dimples (AU14), chin raiser (AU17), lips funneler (AU22), lips puckerer (AU18), lips roll (AU28), and cheek blow (AU33). Nine visemes are also included representing the following phonemes /ah/, /uh/, /axr/, /eh/, /l/, /m/, /n/, /f/, /iy/ and one extreme facial movement: biting their own top lip.

In order to generate a smooth and low-noise 3D mesh from noisy and incomplete depth maps, we aggregated multiple depth maps from different view points in order to construct a full view of the face for each expression and subject by using the Kinect Fusion algorithm[2] [60], [61]. Thus, a 3D facial surface was obtained for each expression of each subject. Figure 3 shows three examples of obtained scans.

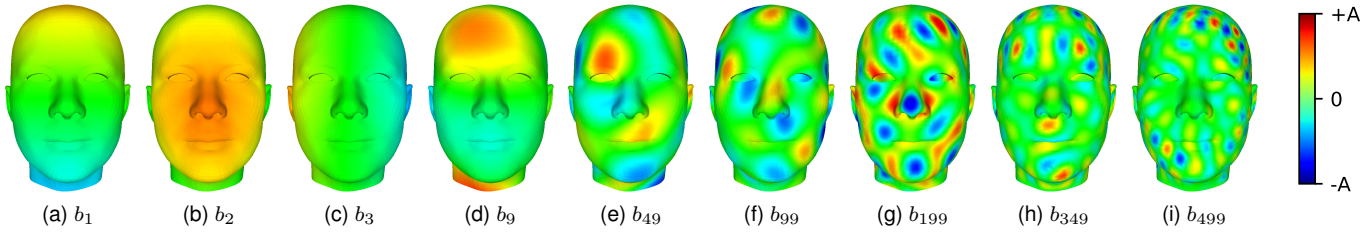
4.1 Nonrigid alignment of the database scans

As mentioned in section 2, in order to allow for statistical modeling of the faces in EPFL3DFace database, these need to be put in dense correspondence. We nonrigidly align all the scans in the database such that all the expressions of all the subjects share a common parametrization using the method described in section 3. This allows for statistical modeling of the variations due both to the identity and the expression.

Our method is based on the deformation of a single template shape. The advantage of not requiring a full statistical shape model (see sec. 2) but only a static template comes at the price of a larger sensitivity to the initialization. This implies that in order to converge to the target, the initial template to be deformed should be close already. Since we observed that the 3D shape of the face varies significantly due to changes in facial expressions, we decided to use a separate template for each facial expression.

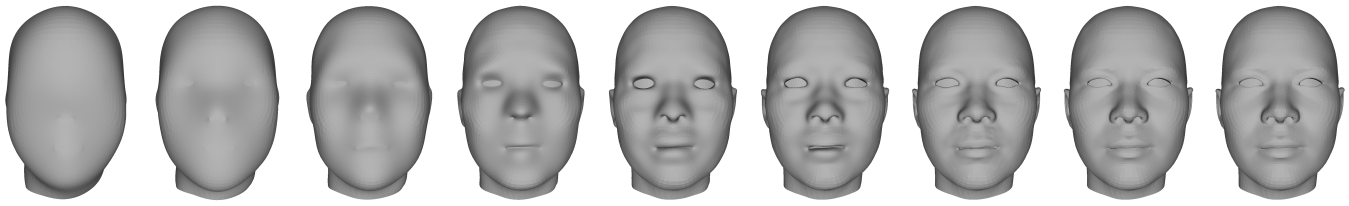
We take advantage of the FaceWarehouse [21] database and compute one mean shape for each expression. For

2. A lightweight, reworked and optimized version of KinFu, originally shared in PCL in 2011, is available on https://github.com/Nerei/kinfu_remake



(a) b_1 (b) b_2 (c) b_3 (d) b_9 (e) b_{49} (f) b_{99} (g) b_{199} (h) b_{349} (i) b_{499}

Fig. 4. Visualization of some of the spectral bases b_i. The amplitude of the deformation for each vertex is normalized over the first 500 bases where $-A$ is the maximum deformation amplitude towards the inside of the surface and $+A$ the maximum towards the outside of the surface.



(a) $N = 50$ $E = 78.28\%$ (b) $N = 100$ $E = 82.13\%$ (c) $N = 200$ $E = 85.66\%$ (d) $N = 500$ $E = 90.02\%$ (e) $N = 1000$ $E = 92.91\%$ (f) $N = 2000$ $E = 95.52\%$ (g) $N = 5000$ $E = 98.74\%$ (h) $N = 10000$ $E = 99.95\%$ (i) original

Fig. 5. (a)-(h) Reconstructions of the FaceWarehouse neutral mean shape using the first N bases, keeping E percent of the energy. (i) The original shape.

each expression in EPFL3DFace, we select as template the FaceWarehouse mean shape closest to that expression. Some expressions have direct correspondences in both databases as the set of expressions from FaceWarehouse is included in EPFL3DFace. For the remaining expressions in EPFL3DFace, we manually selected the closest corresponding expression in FaceWarehouse.

An important advantage of using different templates for each expression is that we do not need to perform any kind of expression transfer. Indeed, the templates of all expressions are already registered together. After registration, the scans of different expressions are in dense correspondence, since the templates used for registration are in dense correspondence.

In practice, we do not evaluate all the vertices of the source in the implicit function of equation (2), but only the vertices lying on the face. Due to the fact that we use the closest expression mean shape of FaceWarehouse as the source for each expression in EPFL3DFace, the topologies of the source and the targets are very different. FaceWarehouse mean shapes are closed surfaces, homeomorphic to a sphere, whereas the scans are bounded surfaces, homeomorphic to a plane. Moreover, the scans of the head are only partial and information is missing on the top and the back of the head. That is not the case with the FaceWarehouse shapes. Trying to align all the vertices of such shapes onto our scans would not be reasonable as they do not share the same topologies and do not contain the same information even though there is an overlap. Thus, we define the set of landmarks lying on the face to use in the implicit distance computation defined in equation (2). Note that the deformation is still applied to the whole shape. In summary, the whole source shape is deformed such that the distance between vertices on the face and the target is minimized.

This nonrigid alignment process is repeated for each scan in the database, resulting in a database of registered 3D surfaces of 120 subjects, performing 35 different expressions and facial movements. This database is available to the research community upon request.

5 Results

In this section we discuss qualitative results obtained with the proposed method on the collected database. In subsection 5.1, we show a few of the manifold harmonic bases that are used to constrain the nonrigid deformation as well as different reconstructions obtained with a varying number of bases and discuss the influence of the number of bases. In subsection 5.2, we then show visual results and compare the obtained deformed shapes with their corresponding targets. We also provide detailed visualization of the spectral deformation process and analyze the evolution of the different terms in the objective function. Finally in subsection 5.3, we visualize the manifolds of shapes and compare these manifolds between an existing database, FaceWarehouse, and our new database.

The lack of ground truth is the main obstacle to a quantitative validation of the method. Indeed, as it is a dense registration problem, the locations of each and every landmarks of the source would need to be manually annotated. Depending on the number of vertices in the source, this represents several thousands of 3D locations for each 3D face scan. Moreover, this problem is largely under-constrained. Ultimately, the topology and geometry of the target is transferred to the source, but these are not uniquely defined by the 3D locations of the vertices. As an example, moving one vertex of the source along the surface of the target does not necessarily change the quality of the alignment.

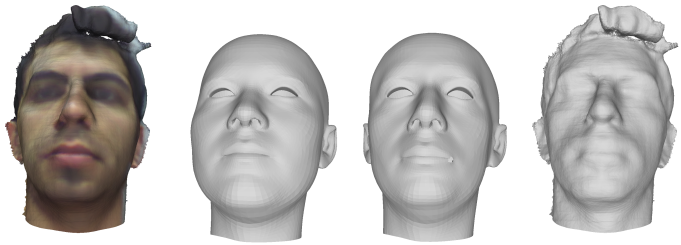


Fig. 6. Alignment results. (a) Color target (b) Rigidly aligned source (c) Result of the nonrigid alignment (d) Target without color (for better comparison).

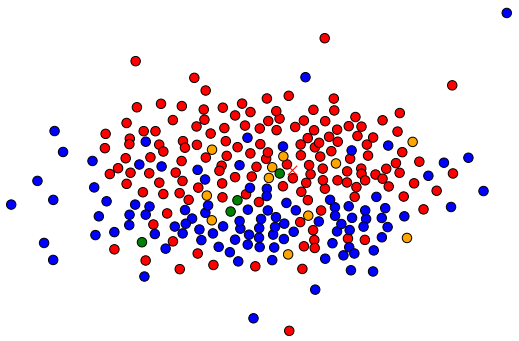


Fig. 7. Database subspaces visualization. ● FaceWarehouse, × Mean shape FaceWarehouse, ● Caucasian subjects from *EPFL3dFace*, ● Asian subjects form *EPFL3dFace*, ● South American subjects from *EPFL3dFace*.

5.1 Spectral basis visualization

Figure 4 shows the first 3 bases and a few other bases corresponding to higher frequencies. Note that basis 0 is constant and is not depicted in the figure. As expected, spectral bases corresponding to lower frequencies show smoother deformations of the surface, whereas higher frequencies provide more localized deformations. The choice of the number of bases to consider in the deformation model is thus guided by the level of details at which the deformation is expected to fit. A very important consideration is that this resolution is only the resolution of the deformation and not the resolution of the obtained mesh. Indeed, the spectral content of all other frequencies outside the frequency band considered in the deformation model is retrieved from the source shape $\bar{\boldsymbol{p}}$ in equation (7).

In order to get a better intuition of the resolution of the deformation, figure 5 shows different reconstructions of the FaceWarehouse [21] mean shape with neutral expression. These were obtained by computing the MHT, transforming the shape into the spectral domain, setting all the spectral coefficients $\tilde{x}$ to zero except the first N coefficients and taking the inverse transform to return to the spatial domain. In short, the source has been filtered with a low-pass filter, whose cut-off frequency varies with the number of bases kept.

Experimentally, we found that keeping the first 500 bases is a reasonable trade-off between the resolution of the deformation and the compactness of the deformation model. The energy of the template that is kept in these 500 bases corresponds to 90.02% of the total energy. With 500 bases, the resolution of the deformation is sufficient to deform shapes with a given expression toward the scans representing the same expression or close ones on subjects with different identities, as explained in section 5.2.

5.2 Spectral alignment

We present the results of the complete nonrigid alignment process on a neutral scan of the EPFL3DFace database in figure 6. Figure 6a shows the clean and normalized color scan from the database. The corresponding mean shape from FaceWarehouse is then rigidly aligned to the scan using 3D landmarks, as detailed in section 3.1. In that case, the source is the neutral expression mean shape. The resulting scaled, translated, and rotated mean shape is shown in figure 6b. That rigidly aligned source is then nonrigidly deformed following the method described in section 3.2 and the result is shown in figure 6c. For better visual comparison, the target is shown again, without texture, in figure 6d. It should be noted that even though the rigid alignment does not retrieve the exact pose of the target, as shown by a comparison of the head poses between figures 6a and 6b, this misalignment is corrected during the nonrigid alignment by the linearized rotation term of the transformation model described in equation (8). Figure 9 presents additional results on two subjects performing seven other expressions or facial movements. Due to the figure's size, it is placed at the end of the paper.

In order to get a better understanding of the spectral deformation process, figure 8 shows the evolution of the different terms in the objective function as well as corresponding shapes, magnitudes of deformation, and distances to the target for a few steps of the optimization. Overall, the data term and the sum of all terms decrease with the number of iterations and seem to have converged at the end of the optimization process. The role of the bending regularization term is clear in the first steps of the optimization, where it prevents extreme, non-realistic deformations to dominate as seen in iteration 1 in figure 8a.

5.3 Facial manifold visualization

In order to validate the intuition that training 3D face models using scans of people from different populations yields different manifolds, we visualize the manifold of scans from the FaceWarehouse database as well as EPFL3DFace using t-SNE [62]. Following the idea of Booth et al. [22], we train a simple principal component analysis (PCA) model of the neutral faces in FaceWarehouse and EPFL3DFace, project the training samples onto that d-dimensional subspace and use t-SNE to generate a 2D visualization of that subspace. We then label the samples according to which database they belong to. Figure 7 shows the resulting visualization.

More specifically, we represent each shape as a vector $S = (x_0, y_0, z_0, ..., x_k, y_k, z_k)$, with $k = 5956$, the number of vertices lying on the face, as defined in section 4.1. We then compute a PCA decomposition of the matrix whose rows

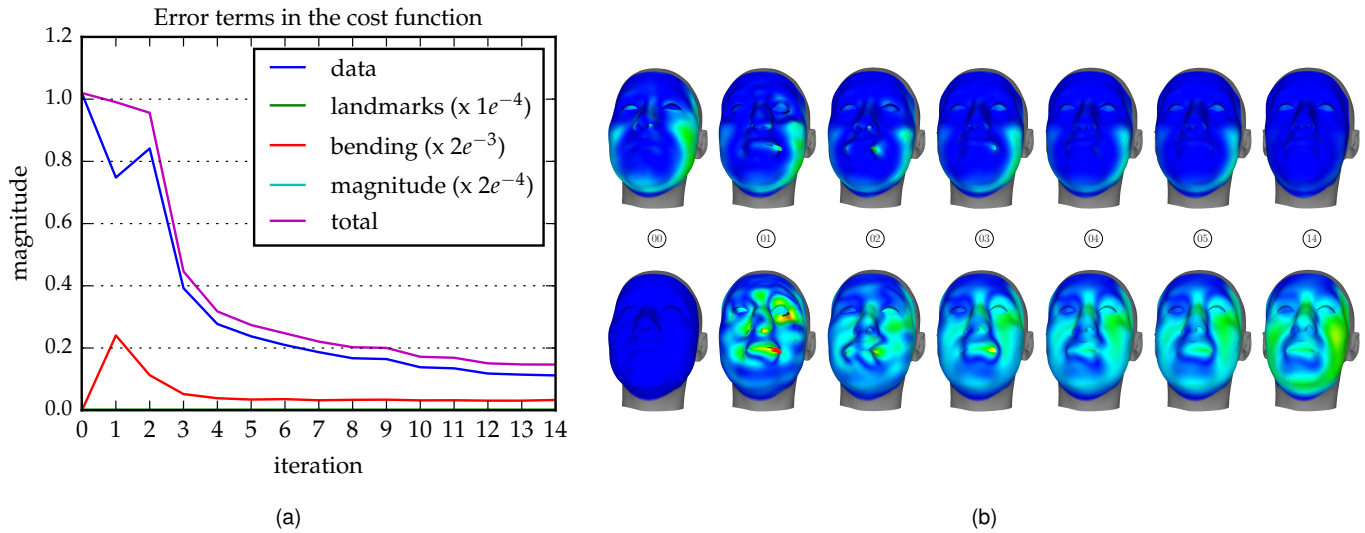


Fig. 8. Evolution of the objective function and corresponding shapes during optimization (a) Evolution of the individual terms in the objective function as well as the total cost for each iteration of the optimization. (b) The first row shows the distance to the target, normalized over the whole sequence and the second row the amplitude of the deformation for different steps of the optimization process, normalized in the same way.

are the shape vectors. Only the first 96 eigenvectors, which together explain more than 99% of the variance of the data, are kept. Each shape is then projected on the PCA basis, thus effectively reducing the original high number of dimensions of these. The new parametrization of the shapes in the PCA basis is the input to the t-SNE algorithm. t-SNE then projects the data to a low-dimensional subspace, typically 2D, while preserving similarities between data points and allows to visualize the structure of the data [62]. In this 2D space, we then label each point, which corresponds to each shape, according to the database that shape belongs to. For EPFL3DFace, we also chose to label the different ethnicities differently. This is not possible for FaceWarehouse, as we do not have the ground truth labels for the ethnicity of the subjects.

In the visualization in figure 7, shapes from different databases appear to be clustered and these clusters span different part of the subspace. Moreover, all the shapes from EPFL3DFace are obtained by nonrigidly deforming the mean shape of the corresponding expression in FaceWarehouse, represented as a cross in figure 7. This mean shape, the neutral expression in that case, is thus effectively deformed in a way that is complementary to the existing shapes in FaceWarehouse.

6 Conclusion, Discussion and Future Work

In this paper, we introduce a new method to nonrigidly register a template to 3D surfaces. We take advantage of spectral geometry processing methods and propose to use manifold harmonic transform (MHT) to constrain the deformation of the template, while enforcing smoothness and reducing the number of parameters in the deformation model. More advanced use of the spectral nature of that deformation model needs to be further investigated. For example, it could be beneficial to select a different frequency band in which to deform the template, depending on the template, the level of details and the application. In our case, we show qualitatively that we obtain a reasonable level of details using only the first 500 spectral bases.

In addition, we propose to use an implicit surface representation based on multilevel partition of unity (MPU) for the target. This presents two main advantages: first, this new representation of the target surface allows to denoise the surface by approximating rather than interpolating it and fill in missing data. Second, the evaluation of the distance to the target is considerably simplified and is reduced to evaluating the implicit function, avoiding the need for correspondences.

Finally, we apply the proposed method on 3D facial scans in order to align them, or put them in dense correspondence. This is required to perform statistical analysis on the set of shapes and ultimately train a 3D statistical shape model. The nonrigidly registered set of shapes constitutes a new database of 3D facial expressions, EPFL3DFace, containing 120 subjects performing 35 different facial expressions and movements. This database is available to the research community upon request. We show that EPFL3DFace is complementary to the existing FaceWarehouse database and that both of them can be combined such that the number of subjects is increased by 80% and that they span a larger subspace.

References

[1] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect Database for Face Recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 11, pp. 1534–1548, 2014.

[2] N. M. Arar, H. Gao, H. K. Ekenel, and L. Akarun, "Selection and combination of local Gabor classifiers for robust face verification," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2012, pp. 297–302.

[3] C. Ding, D. Tao, I. Systems, and I. Technology, "A Comprehensive Survey on Pose-Invariant Face Recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, 2016.

[4] G. Fanelli, J. Gall, and L. Van Gool, "Real Time 3D Head Pose Estimation: Recent Achievements and Future Challenges," in *Proceedings of the International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Rome, 2012.

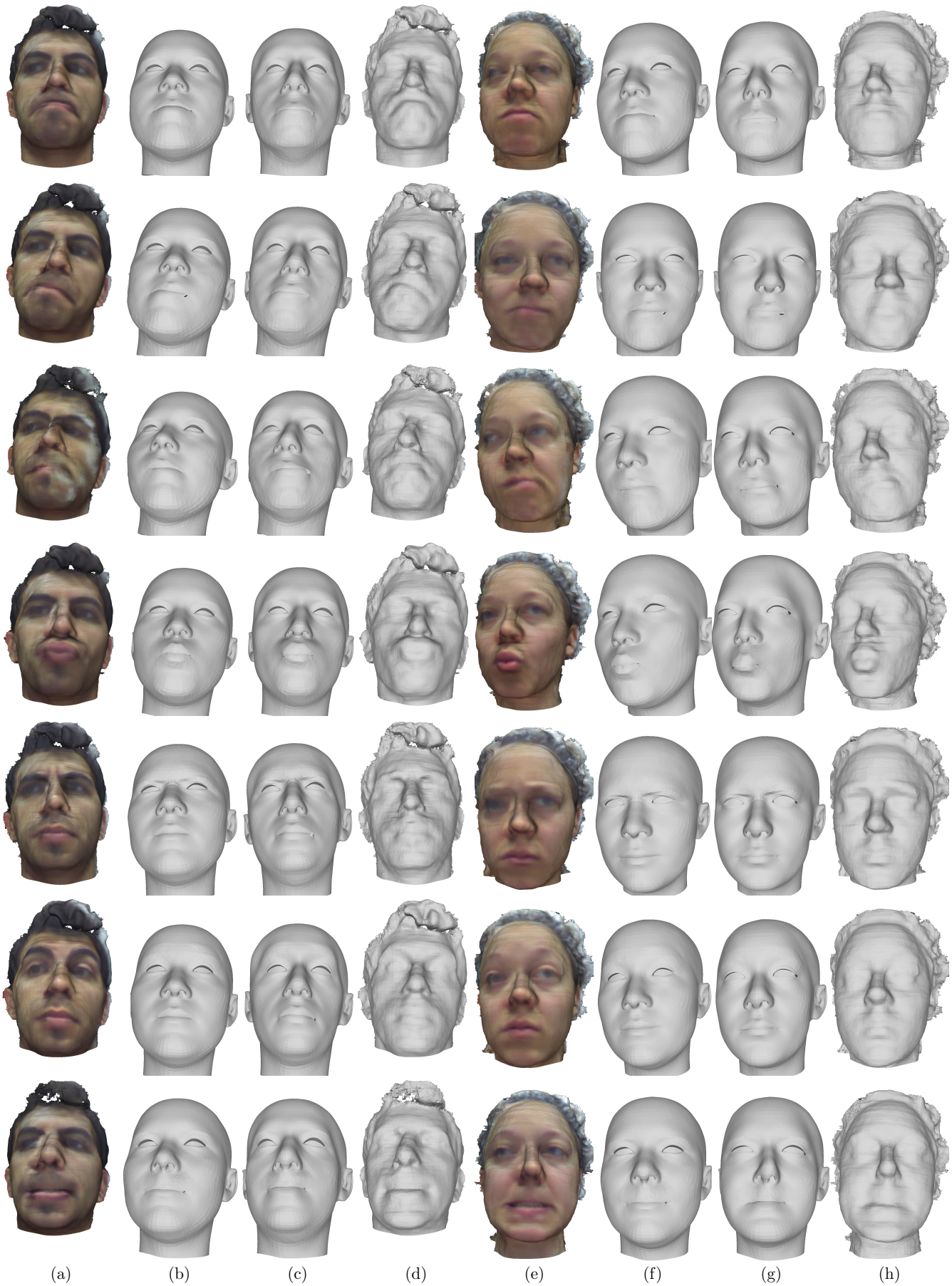


Fig. 9. Alignment results for subject 000 and subject 001 on the following facial movements: jaw forward, jaw left, jaw right, lip puckerer, eyebrows lower, eyebrows raiser and upper lip bite. (a) and (e) Color target. (b) and (f) Rigidly aligned source. (c) and (g) Result of the nonrigid alignment. (d) and (h) Target without color (for better comparison)

[5] K. Alberto, F. Mora, and J.-M. Odobez, "Gaze Estimation from Multimodal Kinect Data," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 4321–4326.

[6] N. M. Arar, Hua Gao, and J.-P. Thiran, "Robust gaze estimation based on adaptive fusion of multiple cameras," in *Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2015, pp. 1–7.

[7] M. Zimmermann, M. Mehdipour Ghazi, H. K. Ekenel, and J.-P. Thiran, "Visual Speech Recognition Using PCA Networks and LSTMs in a Tandem GMM-HMM System," in *Proceedings of Asian Conference on Computer Vision- Workshop Multi-view Lip-reading Challenge (ACCVW)*, 2016.

[8] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, oct 2012.

[9] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "FERA 2015 - second Facial Expression Recognition and Analysis challenge," in *Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 06, 2015, pp. 1–8.

[10] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2016, pp. 1–8.

[11] A. Yüce, N. M. Arar, and J. P. Thiran, "Multiple local curvature gabor binary patterns for facial action recognition," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8212 LNCS, pp. 136–147, 2013.

[12] A. E. Ichim, S. Bouaziz, and M. Pauly, "Dynamic 3D avatar creation from hand-held video input," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 45:1–45:14, jul 2015.

[13] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face / Off : Live Facial Puppetry," in *Eurographics/ACM SIGGRAPH Symposium on Computer Animation*, 2009, pp. 7–16.

[14] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3D Shape Regression for Real-time Facial Animation," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 41.1–41.10, 2013.

[15] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," *ACM Transactions on Graphics, SIGGRAPH 2011*, pp. 1–9, 2011.

[16] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. NieBner, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016, pp. 2387–2395.

[17] N. M. Arar, N. K. Bekmezci, G. Fatma, M. Bilgisayar, and H. K. Ekenel, "Real-time Face Swapping in Video Sequences: Magic Mirror," in *Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, 2012.

[18] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the conference on Computer graphics and interactive techniques (SIGGRAPH)*. New York, USA: ACM Press, 1999, pp. 187–194.

[19] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM Transactions on Graphics*, vol. 24, no. 3, p. 426, jul 2005.

[20] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, sep 2009, pp. 296–301.

[21] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou, "FaceWarehouse: A 3D Facial Expression Database for Visual Computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, mar 2014.

[22] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3D Morphable Model learnt from 10'000 faces," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[23] P. Huber, G. Hu, R. Tena, P. Mortazavian, W. P. Koppen, W. J. Christmas, M. Rätsch, and J. Kittler, "A Multiresolution 3D Morphable Face Model and Fitting Framework," in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and and Technology Publications, 2016, pp. 79–86.

[24] T. Bolkart and S. Wuhrer, "A Groupwise Multilinear Correspondence Optimization for 3D Faces," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 3604–3612.

[25] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, *Numerical Geometry of Non-Rigid Shapes*, springer ed., ser. Monographs in Computer Science. New York, NY: Springer New York, 2008.

[26] O. van Kaick, H. Zhang, G. Hamarneh, and D. Cohen-Or, "A Survey on Shape Correspondence," *Computer Graphics Forum*, vol. 30, no. 6, pp. 1681–1707, sep 2011.

[27] G. K. L. Tam, Zhi-Quan Cheng, Yu-Kun Lai, F. C. Langbein, Yonghuai Liu, D. Marshall, R. R. Martin, Xian-Fang Sun, and P. L. Rosin, "Registration of 3D Point Clouds and Meshes: A Survey from Rigid to Nonrigid," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 7, pp. 1199–1217, jul 2013.

[28] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime Faces: High Resolution Capture for Modeling and Animation," *Proceedings SIGGRAPH*, pp. 548–558, 2004.

[29] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Calculus of nonrigid surfaces for geometry and texture manipulation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 5, pp. 902–913, 2007.

[30] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, and B. Gökberk, "Bosphorus Database for 3D Face Analysis," in *Workshop on Biometrics and Identity Management (BIOID)*, 2008, pp. 47–56.

[31] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. Ieee, 2006, pp. 211–216.

[32] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 80, jul 2007.

[33] T. Vetter and V. Blanz, "Estimating coloured 3d face models from single images: An example based approach," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1407, pp. 499–513, 1998.

[34] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes," *ACM Transactions on Graphics*, vol. 22, no. 3, p. 587, 2003.

[35] I. Mpiperis, S. Malassiotis, and M. Strintzis, "Bilinear Models for 3-D Face and Facial Expression Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 498–511, sep 2008.

[36] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, feb 1992.

[37] B. Amberg, S. Romdhani, and T. Vetter, "Optimal Step Nonrigid ICP Algorithms for Surface Registration," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2007, pp. 1–8.

[38] Shiyang Cheng, I. Marras, S. Zafeiriou, and M. Pantic, "Active nonrigid ICP algorithm," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, may 2015, pp. 1–8.

[39] G. Passalis, I. Kakadiaris, T. Theoharis, G. Toderici, and N. Murtuza, "Evaluation of 3D Face Recognition in the presence of facial expressions: an Annotated Deformable Model approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW)*, 2005, pp. 171–171.

[40] I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis, "Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 640–649, 2007.

[41] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer, "Review of statistical shape spaces for 3D data with comparative analysis for human faces," *Computer Vision and Image Understanding*, vol. 128, pp. 1–17, 2014.

[42] N. J. Mitra, N. Gelfand, H. Pottmann, and L. Guibas, "Registration of point cloud data from a geometric optimization perspective," in *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing - SGP '04*. New York, New York, USA: ACM Press, 2004, p. 22.

[43] H. Li, B. Adams, L. J. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," *ACM Transactions on Graphics*, vol. 28, no. 5, p. 1, 2009.

[44] E. Zell and M. Botsch, "ElastiFace," in *Proceedings of the Symposium*

on Non-Photorealistic Animation and Rendering - NPAR '13. New York, New York, USA: ACM Press, 2013, p. 15.

[45] J. Tena, M. Hamouz, A. Hilton, and J. Illingworth, "A Validated Method for Dense Non-rigid 3D Face Registration," in *2006 IEEE International Conference on Video and Signal Based Surveillance*. IEEE, nov 2006, pp. 81–81.

[46] Zhili Mao, J. Siebert, W. Cockshott, and A. Ayoub, "Constructing dense correspondences to analyze 3D facial change," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* IEEE, 2004, pp. 144–148 Vol.3.

[47] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching," *Proceedings of the National Academy of Sciences*, vol. 103, no. 5, pp. 1168–1172, jan 2006.

[48] ——, "Expression-Invariant Representations of Faces," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 188–197, jan 2007.

[49] J. Huang, X. Shi, X. Liu, K. Zhou, L.-Y. Wei, S.-H. Teng, H. Bao, B. Guo, and H.-Y. Shum, "Subspace gradient domain mesh deformation," *ACM Transactions on Graphics*, vol. 25, no. 3, p. 1126, 2006.

[50] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Transactions on Graphics*, vol. 23, no. 3, p. 399, 2004.

[51] B. Vallet and B. Lévy, "Spectral Geometry Processing with Manifold Harmonics," *Computer Graphics Forum*, vol. 27, no. 2, pp. 251–260, apr 2008.

[52] Y. Ohtake, A. Belyaev, M. Alexa, G. Turk, and H.-P. Seidel, "Multi-level partition of unity implicits," *ACM SIGGRAPH 2003 Papers on - SIGGRAPH '03*, p. 463, 2003.

[53] C. Qu, H. Gao, E. Monari, J. Beyerer, and J.-P. Thiran, "Towards robust cascaded regression for face alignment in the wild," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW)*. IEEE, jun 2015, pp. 1–9.

[54] X. Xiong and F. D. Torre, "Supervised Descent Method for Solving Nonlinear Least Squares Problems in Computer Vision," in *arXiv preprint arXiv:1405.0601*, 2014, pp. 1–15.

[55] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-Squares Fitting of Two 3-D Point Sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5, pp. 698–700, sep 1987.

[56] O. Sorkine, "Least-squares rigid motion using svd," Tech. Rep. February, 2009.

[57] G. Taubin, "A signal processing approach to fair surface design," in *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. New York, New York, USA: ACM Press, 1995, pp. 351–358.

[58] B. Lévy and H. Zhang, "Spectral Mesh Processing," in *ACM SIGGRAPH Course Notes*, 2010.

[59] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel, "A statistical method for robust 3D surface reconstruction from sparse data," in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.* IEEE, 2004, pp. 293–300.

[60] R. a. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. IEEE, oct 2011, pp. 127–136.

[61] S. Izadi, A. Davison, A. Fitzgibbon, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, and D. Freeman, "KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moveing Depth Camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. New York, New York, USA: ACM Press, 2011, pp. 559–568.

[62] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

Gabriel L. Cuendet received his B.Sc. and M.Sc. degrees in electrical engineering with specialization in biomedical engineering from the Ecole Polytechnique Fedérale de Lausanne, Switzerland, in 2012, where he is currently working toward the Ph.D. degree in developing facial image analysis for medical diagnosis applications. His research is focused on computer vision methods for 2D and 3D facial landmarks detection and tracking.

Christophe Ecabert has received his B.Sc. degree in electrical engineering from the Haute Ecole Spécialisée de Suisse occidentale in 2010 and M.Sc. degrees in electrical engineering from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, in 2014. He is currently pursuing a PhD degree at the Signal Processing Laboratory (LTS5) at EPFL. His research aims at developing 3D based methods for facial image analysis.

Marina Zimmermann received the B.Sc. degree from Jacobs University Bremen, Bremen, Germany, in 2011 and the M.Sc. degree in electrical engineering from Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland in 2013, where she is currently working towards a Ph.D. degree at the Signal Processing Laboratory. Her research interests include facial image analysis, with a focus on visual speech recognition.

Hazım K. Ekenel is an Associate Professor at the Department of Computer Engineering in Istanbul Technical University in Turkey. He received his Ph.D. degree in Computer Science from the University of Karlsruhe (TH) in 2009. He has founded the Facial Image Processing and Analysis group at the Department of Computer Science in Karlsruhe Institute of Technology. He has received the EBF European Biometric Research Award in 2008 for his contributions to the field of face recognition and with the systems they have developed with his team, he received the Best Demo Award at the IEEE International Conference on Automatic Face and Gesture Recognition in 2008. He has coordinated the Benchmarking Facial Image Analysis Technologies (BeFIT) initiative and he was the primary organizer of the BeFIT workshops 2011 and 2012.

Jean-Philippe Thiran is Associate Professor of Image Processing and director of the Signal Processing Laboratory (LTS5) at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He also holds an Associate Professor position with the Department of Radiology of the University Hospital Center (CHUV) and University of Lausanne (UNIL). His research field is image analysis and multimodal signal/image processing, with applications in many domains including medical image analysis, human-computer interaction, remote sensing of the Earth, and surveillance. Dr Thiran is author or co-author of more than 130 journal papers, 9 book chapters, more than 185 papers in peer-reviewed proceedings of international conferences, and holds 4 international patents. He is currently an associate editor of the IEEE Transactions on Image Processing and a reviewer for many journals and conferences. He is a senior member of the IEEE.