Dictionary Learning based 3D Morphable Model Construction for Face Recognition with Varying Expression and Pose

Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, Alberto Del Bimbo
Media Integration and Communication Center
University of Florence, Florence, Italy
{claudio.ferrari, giuseppe.lisanti, stefano.berretti, alberto.delbimbo}@unifi.it

Abstract

In this paper, we propose a new approach for constructing a 3D morphable model (3DMM) and experiment its application to face recognition. Differently from existing solutions, the proposed 3DMM is constructed from a training set that includes a large spectrum of variability in terms of ethnicity and facial expressions. By exploiting annotated landmarks available in the training data, we are able of establishing dense correspondence across training scans also in the presence of strong facial expressions. The 3DMM is then constructed by learning a dictionary of basis components, instead of using the traditional approach based on PCA decomposition. Finally, we cast the proposed dictionary learning DL-3DMM to a rigid / non-rigid deformation framework, which includes pose estimation and regularized ridge-regression fitting to 2D images. Comparative results between the DL-3DMM and its PCA counterpart are reported, together with face recognition results for images with large pose and expression variations.

1. Introduction

The idea of characterizing the statistical variability of the traits of the human face dates back to the early '80s with the work of Farkas [11]. In his studies, the face variability was modeled using anthropometric measures between a set of facial landmarks. More recently, Sirovich and Kirby [24] shown that Principal Component Analysis (PCA) could be used on a collection of training face images to form a set of basis features, known as *eigenpictures*, that can be linearly combined to reconstruct images in the original training set. This idea was extended further in the work of Blanz and Vetter [4], that proposed to construct a 3D morphable model (3DMM) from a set of example 3D face scans. Similarly to the 2D case, the idea here is to capture the 3D face variability in the training data using an average model and a set of deformation components learned using PCA decomposition. The statistical model is then capable of generating new face instances with plausible shape and appearance. Since that work, 3DMM variants have been used in computer graphics for face inverse lighting [23, 27] and reanimation [3], 3D shape estimation from 2D image face data [28], pose robust face recognition [5, 16], 3D face recognition [1], and several others tasks. In particular, there is the feeling that the statistical information brought by the 3DMM can be exploited to improve face recognition performance in the case of 2D images acquired in the real, with large pose variations, occlusions, illumination changes, and facial expressions (face recognition in the "wild"). The idea here is that given a single face image under unknown pose and illumination, the 3DMM can solve its 3D shape, texture, pose and illumination parameters simultaneously, using Gauss-Newton optimization [21] or regression [28] to minimize the difference between the synthetic image rendered by the 3DMM and the input image. The performance of such an approach are bounded by the specific characteristics of the 3DMM which, in turn, depend on its construction and fitting. More specifically, constructing a 3DMM for face recognition applications requires the following aspects be considered: *i)* A training data set should be acquired at good resolution, and it should include a significant sample of the human face variability in terms of gender, ethnicity, age, and facial expressions; moreover, a dense correspondence is required between the 3D facial scans in the training set; *ii)* The statistical variability of the training scans is captured in a compact form using a statistical modeling (3DMM), capable of generating new face instances; *iii)* Defining an appropriate fitting approach, which can deform the 3DMM adapting it to 2D target images.

In this paper, we will address the above aspects, by proposing original solutions for constructing and using a 3DMM for face recognition from 2D images.

1.1. Related work

In their seminal work, Blanz and Vetter [4] at the Max Plank Institute (MPI) were the first to present a complete

978-1-4673-8332-5/15 $31.00 © 2015 IEEE
DOI 10.1109/3DV.2015.63

solution to derive a 3DMM by transforming the shape and texture from an example set of 3D face scans into a vector space representation based on PCA. However, the variability in the training dataset they used is not very large (200 scans of Caucasian young individuals were included), reducing the capability of the model of generalizing to different ethnicity. The optical flow algorithm used to establish dense correspondence in the training data also limits the applicability to faces with facial expressions. Furthermore, being based on the linear combination of Principal Components (PC), holistic deformations are obtained, where each PC acts on every vertex of the model. To account for this effect, the face is divided into four disjoint subregions (i.e., eyes, nose, mouth and a surrounding region) that are morphed independently. A complete 3D face is finally generated by computing linear combinations for each segment separately and blending them at the borders. Despite of these limitations, the MPI 3DMM has proved its effectiveness in several applications, inspiring most of the subsequent work. In [5], Blanz and Vetter used the 3DMM to simulate the process of image formation in 3D space, using computer graphics, and estimated 3D shape and texture of faces from single images. The estimate is achieved by fitting the MPI 3DMM to 2D images. This allows for face recognition across variations in pose, and across a wide range of illuminations. Romdhani and Vetter [21], used the MPI 3DMM for face recognition by enhancing the deformation algorithm through the inclusion of various image features, such as the edges or the location of the specular highlights. The 3D shape, texture and imaging parameters are then estimated by maximizing the posterior of the parameters given these image features.

The MPI 3DMM was further refined into the Basel Face Model (BFM) by Paysan *et al*. [19]. This improves on previous models by offering higher shape and texture accuracy thanks to a better scanning device (though the training subjects are yet 200 Caucasians only), and less correspondence artifacts thanks to an improved registration based on the optimal non-rigid iterative closest point (ICP) algorithm [2]. However, since the optimal non-rigid ICP cannot handle large missing regions and topological variations, expression variations are not accounted for in the training data also in this case. In the work of Amberg *et al*. [1], a 3DMM was constructed by a set of 270 neutral plus 135 expressive scans for the purpose of 3D face recognition. Since a modification of the non-rigid ICP algorithm [2] was used to perform dense alignment in the training data, and also to fit the 3DMM to target scans, the method may fail in the case of large missing regions and topological variations. Bustard and Nixon [9] modified the 3DMM by including the ear region, and proposed it for ear and face recognition.

Patel and Smith [18], instead, have shown that the statistical tools of thin-plate splines and Procrustes analysis can be used to construct a 3DMM. In particular, Procrustes analysis is used to establish correspondence between a set of manually labelled landmarks of the face. The averages of these landmarks are then used as anchor points to construct a complete deformable model by interpolating the regions between landmarks using thin-plate splines. A statistical model for 3D human faces in varying expression has been proposed by Brunton *et al*. [7], which decomposes the face using a wavelet transform, and learns many localized, decorrelated multilinear models on the resulting coefcients. In the work of Kakadiaris *et al*. [12], a model based technique is used for 3D face recognition. This is based on the deformation of an Active Face Model (AFM), which is obtained as the average facial 3D mesh from statistical data. However the model has low resolution and synthetic appearance. An alternative solution to fit the 3DMM is presented by Zhu *et al*. [28], which is based on a novel discriminative method for estimating 3D shape from a single image with a 3DMM. They proposed to estimate the shape parameters by learning a regressor, instead of minimizing the appearance difference. Compared with the traditional analysis-by-synthesis framework, the discriminative approach makes it possible to utilize large databases to train a fitting model.

For a comprehensive review on statistical 3D face modeling, we refer the reader to the work of Brunton *et al*. [8].

1.2. Our contribution and paper organization

Motivated by the above considerations, in this paper, we propose a new method for constructing a 3DMM, and apply it in the context of face recognition. Initially, we point out the importance of selecting a representative training set of 3D faces for the model construction, where a large spectrum of face variations is included. Then, as first contribution of this work, we propose an original method to obtain a dense vertex-by-vertex correspondence across the training data. With respect to existing solutions, the proposed method has the advantage of not depending on the choice of a reference model to transfer the correspondence, and thus can act independently on each scan. The second main contribution consists in the definition of a different approach for the 3DMM construction. In contrast to the PCA decomposition used in most of the state of the art solutions, we learn a dictionary (basis) from the vector field of the deviations between each 3D scan and an average model. Such an average model can be easily derived from a training set, after the dense alignment process, by averaging the vertex positions overall the scans. The proposed framework has been evaluated in two sets of experiments: first, we compared the 2D reprojection error and the 3D reconstruction error obtained with the proposed 3DMM with respect to standard PCA decomposition; then, we experimented the effectiveness of the 3DMM in supporting face recognition from images with large pose variations and expression changes.

The rest of the paper is organized as follows: In Sect. 2, we propose a novel solution for establishing dense correspondence between the faces of a training set with a large spectrum of face variations; In Sect. 3, the 3DMM construction from the training set is then expounded as a dictionary learning process; In Sect. 4, the deformation process used to fit the 3DMM to 2D target images is modeled through pose estimation plus the solution of a ridge-regression problem. A comparative evaluation of the proposed 3DMM with respect to standard PCA decomposition is reported in Sect. 5, together with the application of the 3DMM to face recognition. Finally, discussion and future work are reported in Sect. 6.

2. Database selection and processing

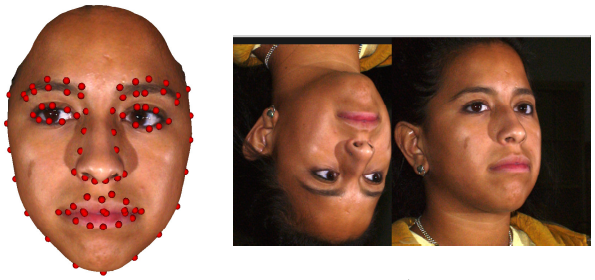
Constructing a 3DMM requires two preliminary steps: selection of 3D scans with the appropriate characteristics to be used as training set; establishing a dense vertex-by-vertex correspondence across the training data.

2.1. Training data

In order to guarantee the 3DMM capability of generalizing to new unseen identities, the training set should include the necessary variability in terms of gender, age and ethnicity. Furthermore, including in the training set scans with facial expressions is also required to enable the 3DMM to generalize to expressive data.

The MPI 3DMM [4], and the most recent BFM [19] have proved their usefulness in computer graphics and recognition contexts. However, both of them are constructed from a training set with limited variability, which includes face scans of 100 female and 100 male subjects, most of them Caucasians, with age between 8 and 62 years (average of 25 years). In both the cases, authors explicitly reported that expressive scans were not included in the training set.

Since the 3DMM aims to capture the human face variability in a consistent way, we propose to use the Binghamton University 3D Facial Expression dataset (BU-3DFE) [26] as training set in the construction of the 3DMM. This dataset is largely used as 3D benchmark for 3D facial expression recognition research, being it publicly available and including a well balanced representation of the human face in terms of gender, age, and ethnicity, as well as of facial expressions. The BU-3DFE contains scans of 44 females and 56 males, with ranging age from 18 to 70 years old, acquired in a neutral plus six different expressions, namely, *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. Apart of the neutral expression, all the other facial expressions have been acquired at four levels of intensity, from low to exaggerated (2500 scans in total). The subjects are well distributed across different ethnic groups or racial ancestries, including *White*, *Black*, *Indian*, *East-Asian*, *Middle-East Asian*, and *Hispanic-Latino*. Each 3D facial scan is also cropped and associated with a set of 83 manually annotated landmarks located in correspondence to the most distinguishing traits of the face (see Fig. 1(a)).

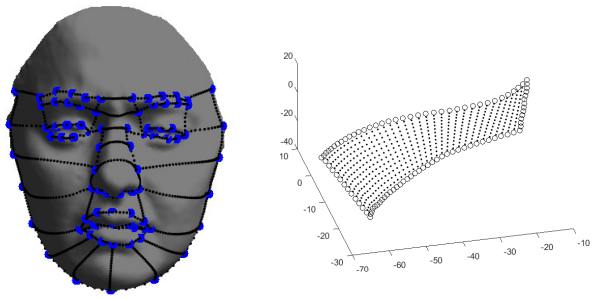


(a) (b)

Figure 1. BU-3DFE: (a) The 83 facial landmarks evidenced on a textured 3D face scan; (b) 2D image captured by the scanner contextually to the 3D scan.

2.2. Establishing 3D dense correspondence

In order to derive a statistics of the face variability in the training data, a dense point-to-point correspondence between the vertices of the training scans should be established. This process can be seen as a sort of mesh reparametrization, where corresponding points in all the scans must have the same semantic meaning (for example, the vertex with index ith must represent the left mouth corner in all the scans). In general, this problem has a not easy solution due to the presence of a limited number of points of the face that are detectable with sufficient accuracy, while large regions are instead characterized by neglectable shape and photometric variations. The presence of facial expressions further increases the difficulty of the problem, especially due to self-occlusions and possible changes in the topology of the surface (as in the case of *mouth-close* / *mouth-open*). In the MPI model [4], a gradient-based optic flow algorithm is modified to establish correspondence between a pair of 3D scans taking into account for color and shape values simultaneously. On facial regions with little structure in texture and shape, such as forehead and cheeks, a smooth interpolation is required to resolve spurious results given by the optic flow algorithm. In [1], a 3DMM was constructed by a set of 270 neutral plus 135 expressive scans. The data were registered with a modification of the non-rigid ICP algorithm proposed in [2]. Though this algorithm has shown its good performance also in the construction of the BFM [19], it has difficulty in handling large missing regions and topological variations. In addition, both the optical flow and the non-rigid ICP methods are applied by transferring the vertex index from a reference model to all the scans. As a consequence, the choice of the reference face can affect the quality of detected correspondences and ultimately of the final model. Moreover,



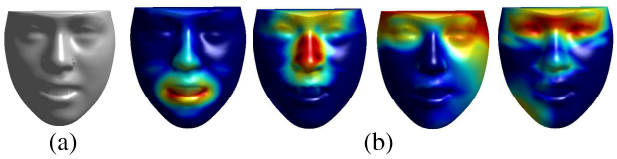
(a) (b)

Figure 2. (a) A face scan of the BU-3DFE with evidenced the 83 landmarks, and the geodesic paths connecting some of them which are used to partition the face; (b) Geodesic contour of the right *cheek / zygoma* region obtained by connecting a set of landmarks. The geodesic path of the boundary is resampled, so that points on it are at the same geodesic distance each other. The interior of the region is also resampled by connecting, through sampled linear paths, corresponding points on opposite sides of the boundary.

the resulting 3DMM is not uniquely defined. In order to solve these problems, Patel and Smith [18] proposed an alternative approach based on the manual annotations of 104 facial landmarks located on the eyebrows and nose contour, eyes and mouth boundary, and face boundary including ears. These points are used to establish a correspondence, so that the mean coordinates of each landmark can be found. The landmarks of each sample are then warped to the mean landmarks, and thin-plate spline interpolation is applied to this warp. Finally, consistent resampling is performed across all faces, but using the estimated surface between landmarks rather than the real one. Working on 3D sequences, Bolkart and Wuhrer [6] presented an approach to fully automatically register 3D faces in motion. This method predicts landmarks for 3D facial motion sequences by a trained multilinear model, and uses these landmarks to initialize sequence registration. A registered version of the BU-3DFE is used to learn the model, as provided in the work of Salazar *et al*. [22].

In this work, we use the BU-3DFE as training set and propose a new approach to establish a dense point-to-point correspondence between the vertices of the scans, which offers potentially more stable performance. First, geodesic paths between facial landmarks are used to partition the face into a set of non-overlapping regions (this shares some common ideas with the method of Lu and Jain [14]). Then, the interior surface of each region is resampled using points on the geodesic boundary as starting and ending points of geodesic line between them. Details are given below.

Face partitioning – The 83 landmarks manually annotated and released with the BU-3DFE are assumed to be



(a) (b)

Figure 3. (a) The average 3DMM obtained from the scans of the BU-3DFE using the proposed dense correspondence approach; (b) Model deformations obtained applying some basis components. The superimposed heatmap represents the magnitude of the deformation (red = high, blue = no deformation). The localized effect on the face produced by the basis components can be appreciated.

correctly identified and thus are used to provide correspondence between salient points of the training data. Here, we develop on the idea that connecting selected pairs of landmarks through geodesic paths on the surface, it is possible to partition the face into a set of regions, as evidenced in Fig. 2(a). With this approach, 10 regions are identified in each side of the face (spanning the eyebrow, eye, check, jaw and chin), plus 8 regions covering the middle part of the face (including the lips, the region between the upper lip and the nose, the nose, and the region between the eyes). In this way, each face is partitioned into a total of 28 non-overlapping regions, each delimited by a closed geodesic contour passing through a set of landmarks. Interestingly, these regions have a great correspondence with optimal regions obtained by automatically segmenting the face, as reported by De Smet and Van Gool [10]. For computing the geodesic path between two landmarks on the surface, we used the variant of the *Fast Marching* algorithm applied to triangular mesh manifolds [13]. Each geodesic path is then resampled with a predefined number of points at equal geodesic distance. As an example, Fig. 2(b) shows with circles the 3D plot of the resampled *geodesic contour* which delimits the right *cheek / zygoma* region comprised between the face boundary and the nose.

Region sampling – Using the above partitioning of the face, we can sample the interior surface of the regions in a consistent way, so that points of homologous regions are in dense correspondence across all the training scans. This is obtained by using the geodesic contour of the region to guide the dense resampling of its interior surface. The idea here is to connect pairs of sampling points on opposite side of a geodesic contour with a straight geodesic line. This line is then sampled at the desired resolution, as illustrated in Fig. 2(b), where the interior of a region after the sampling is reported. It is interesting to note that being based on the annotated landmarks and their connections, the above approach is also robust to facial expressions. In particular, the presence of landmarks on the mouth, which delimit the internal and external border of the lips, makes it possible to

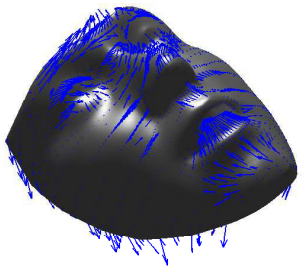


Figure 4. The average model computed on the BU-3DFE. The blue arrows show the field of deviation vectors between the average model and a sample scan of the training set. The dictionary is learnt over such vectors.

maintain regions correspondence also across faces characterized by expressions with *mouth-close* / *mouth-open*. By applying this approach to the scans of the BU-3DFE we obtained the average model of Fig. 3(a).

3. Dictionary learning for 3DMM

Once a dense correspondence is established across the training data, the statistical variability of every vertex of the scans can be modeled. The usual approach to do this is based on PCA decomposition [4]. PCA reduces the space spanned by the training data to an average model and a set of PC corresponding to the eigenvectors of the covariance matrix. Differently, here we propose to learn a *dictionary* of deformation components; to do this, we first learn a basis (*dictionary*) from the training set, then we use linear combinations of such basis to deform the average model. This solution shares some idea with the work of Neumann *et al*. [17], where sparse PCA along with a group sparsity constraint are used to learn a sparse and localized set of deformation components, specifically aimed at mesh animation and 3D modeling tasks.

In general, dictionary learning techniques aim at finding a dictionary $\mathbf{D}$ of k basis vectors (*atoms*), whose linear combination best describes (in this case in terms of reconstruction error) each of the n vectors in the training set. Our goal here is instead estimating the dictionary of basis vectors from the vector field of the deviations between each scan and the average model, so as to reconstruct arbitrary new faces as the sum of the average model and a linear combination of the basis. Dictionary learning is performed by exploiting the training set of 3D face scans in dense correspondence, as obtained in Sect. 2.2. Each training face has m vertices and is represented as a column vector $\mathbf{f}_i \in \mathbb{R}^{3m}$, whose elements are the x, y, z components of all the vertices, that is: $\mathbf{f}_i = [x_1^{(i)}, y_1^{(i)}, z_1^{(i)}, x_2^{(i)}, y_2^{(i)}, z_2^{(i)}, \ldots, x_m^{(i)}, y_m^{(i)}, z_m^{(i)}]$. The deviation field is computed for each model $\mathbf{f}_i$ by subtracting the average one $\mathbf{m} \in \mathbb{R}^{3m}$ computed on the training set:

$$\mathbf{m} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{f}_j \ , \qquad \mathbf{f}_i \leftarrow \mathbf{f}_i - \mathbf{m} \ . \tag{1}$$

A matrix $\mathbf{F} \in \mathbb{R}^{3m \times n}$ is then constructed, having such vectors $\mathbf{f}_i$ as columns: $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]$.

The dictionary learning can be seen as a problem of optimizing the empirical cost function:

$$e_n(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{f}_i, \mathbf{D}) \ , \tag{2}$$

where $\mathbf{D} \in \mathbb{R}^{3m \times k}$ is the dictionary, each column representing a basis vector, and ℓ is a loss function such that $\ell(\mathbf{f}_i, \mathbf{D})$ should be small if $\mathbf{D}$ is "good" at representing the signal $\mathbf{f}_i$. A common expedient to prevent $\mathbf{D}$ from having arbitrarily large values consists in constraining its columns $\mathbf{d}_1, \ldots, \mathbf{d}_k$ to have an ℓ_2-norm less than or equal to one. We will call $\mathcal{C}$ the convex set of matrices verifying this constraint:

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{3m \times k} \ \ s.t.\ \forall j = 1, \ldots, k,\ \mathbf{d}_j^T \mathbf{d}_j \leq 1\} \ . \tag{3}$$

In so doing, we circumscribe the search space of the possible dictionaries, and we estimate the optimal one using a *Elastic-Net* formulation. The Elastic-Net is a type of regression method that linearly combines the sparsity-inducing ℓ_1 penalty and the ℓ_2 regularization. The ℓ_1 norm is known to act as a shrinkage operator, reducing the number of non-zero elements of the dictionary, while the ℓ_2 norm avoids uncontrolled growth of the elements magnitude. By defining $\ell_{1,2}(\mathbf{w_i}) = \lambda_1 \left\|\mathbf{w}_i\right\|_1 + \lambda_2 \left\|\mathbf{w}_i\right\|_2$, where λ_1 and λ_2 are respectively the sparsity and regularization parameters, we can formulate the problem as:

$$\min_{\mathbf{w}_i \in \mathbb{R}^k,\ \mathbf{D} \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \left(\left\|\mathbf{f}_i - \mathbf{D}\mathbf{w}_i\right\|_2^2 + \ell_{1,2}(\mathbf{w_i}) \right) \ . \tag{4}$$

Both penalties are important in this context, since it would be desirable for the dictionary to be (kind of) sparse, so that each component affects only part of the whole model and the elements of the dictionary should be bounded in order to avoid extreme deformations. The above minimization problem is not convex with respect to $\mathbf{D}$. It can be rewritten as a joint optimization problem with respect to the dictionary $\mathbf{D}$ and the coefficients $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n] \in \mathbb{R}^{k \times n}$ of the decomposition, which is convex with respect to each of the two variables $\mathbf{D}$ and $\mathbf{W}$, when the other one is fixed. A natural approach for solving this problem is to alternate between the two variables, minimizing over one while keeping the other one fixed. To this end, we exploited the implementation of the *Online Dictionary Learning for Sparse Coding* presented by Mairal *et al*. [15].

The average model $\mathbf{m}$ and the dictionary $\mathbf{D}$ constitute our 3DMM (DL-3DMM for short).

4. Fitting the 3DMM

In order to fit the 3DMM to a target face in a test image, we need firstly to get an estimate of the 3D pose of the face (rigid transformation), and then to deform the 3DMM to the image (non-rigid transformation).

Pose estimation – Given a face framed in an image, its 3D pose can be estimated by establishing a correspondence between a set of facial landmarks detected both in 2D and 3D. To this end, we employ the 2D face landmark detector defined in [25] that provides sufficient precision in the image localization of 49 facial landmarks. The same set of landmarks is manually annotated once on the vertices of the average model $\mathbf{m}$ in our 3DMM (the indices of these vertices will be indicated with $\mathbf{I}_v$ in the following). In this way, we can indicate with $\mathbf{L} = \hat{\mathbf{m}}(\mathbf{I}_v) \in \mathbb{R}^{3\times|\mathbf{I}_v|}$ the matrix of the coordinates of the 3D landmarks (note that we indicate with $\hat{\mathbf{m}}$ the average model represented in $\mathbb{R}^{3\times m}$ rather than in $\mathbb{R}^{3m}$), while $\mathbf{l} \in \mathbb{R}^{2\times|\mathbf{I}_v|}$ are the landmarks detected in the 2D image. Under an affine camera model [16], the relation between $\mathbf{L}$ and $\mathbf{l}$ is:

$$\mathbf{l} = \mathbf{A}\cdot\mathbf{L} + \mathbf{T}\ , \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{2\times3}$ contains the affine camera parameters, and $\mathbf{T} \in \mathbb{R}^{2\times|\mathbf{I}_v|}$ is the translation on the image. To recover these parameters, firstly we subtract the mean from each set of points, then we recover the affine matrix in a least square sense as $\mathbf{A} = \mathbf{l}\cdot\mathbf{L}^{+}$, where $\mathbf{L}^{+}$ is the pseudo-inverse matrix of $\mathbf{L}$. Then, we estimate the translation as $\mathbf{T} = \mathbf{l} - \mathbf{A}\cdot\mathbf{L}$. Furthermore, using QR decomposition the matrix $\mathbf{A}$ can be decomposed as $\mathbf{A} = \mathbf{S}\cdot\mathbf{R}$, where $\mathbf{S} \in \mathbb{R}^{2\times3}$ expresses the scale parameters along with the shear, and $\mathbf{R} \in \mathbb{R}^{3\times3}$ contains the 3D rotation parameters of the model with respect to the image. Thus, the affine camera model is given by:

$$\mathbf{l} = \mathbf{S}\cdot\mathbf{R}\cdot\mathbf{L} + \mathbf{T}\ . \tag{6}$$

Considering Eq. (6), it is possible to get an estimate of the pose $\mathbf{P}$ as $[\mathbf{S}\cdot\mathbf{R}, \mathbf{T}]$, that permits us to map each vertex of the 3DMM onto the image.

3DMM fitting – Given a target image and the dictionary $\mathbf{D}$, we want to find the pose $\mathbf{P}$, and the combination of the deformation components according to coefficients $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k] \in \mathbb{R}^k$, which minimize a *penalty function* computed on the corresponding landmarks in 2D and 3D. Stated differently, we aim at finding the coding on the learned dictionary $\mathbf{D}$ that (non-rigidly) transforms the average model $\mathbf{m}$, so that its projection with $\mathbf{P}$ (rigid transformation) minimizes the error in correspondence to the landmarks. The coding is formulated as the solution of a regularized *Ridge-Regression* problem:

$$\min_{\mathbf{P},\,\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\cdot\left(\hat{\mathbf{m}}(\mathbf{I}_v) + \sum_{i=1}^{k}\hat{\mathbf{D}}_i(\mathbf{I}_v)\alpha_i\right)\right\|_2^2 + \lambda\left\|\mathbf{w}^{-1}\boldsymbol{\alpha}\right\|_2\ , \tag{7}$$

where λ is a regularization parameter, and $\hat{\mathbf{D}}_i$ indicates a basis component represented in $\mathbb{R}^{3\times m}$ rather than in $\mathbb{R}^{3m}$. We solve this problem by alternating between pose and coefficients estimation. First, we estimate the pose $\mathbf{P}$ and solve for the coefficients $\boldsymbol{\alpha}$:

$$\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\cdot\hat{\mathbf{m}}(\mathbf{I}_v) - \sum_{i=1}^{k}\mathbf{P}\cdot\hat{\mathbf{D}}_i(\mathbf{I}_v)\alpha_i\right\|_2^2 + \lambda\left\|\mathbf{w}^{-1}\boldsymbol{\alpha}\right\|_2\ . \tag{8}$$

Since the pose $\mathbf{P}$, the dictionary $\hat{\mathbf{D}}$, the landmarks $\mathbf{l}$, and $\hat{\mathbf{m}}(\mathbf{I}_v)$ are known, we can define $\mathbf{X} = \mathbf{l} - \mathbf{P}\cdot\hat{\mathbf{m}}(\mathbf{I}_v)$ and $\mathbf{Y} = \mathbf{P}\cdot\hat{\mathbf{D}}_i(\mathbf{I}_v)$, and rewrite Eq. (8) as:

$$\min_{\boldsymbol{\alpha}} \left\| \mathbf{X} - \sum_{i=1}^{k}\alpha_i\mathbf{Y}\right\|_2^2 + \lambda\left\|\mathbf{w}^{-1}\boldsymbol{\alpha}\right\|_2\ , \tag{9}$$

which is analytically solved as a *Ridge-Regression*, where each coefficient is weighted by the inverse of $\mathbf{w}$. These weights provide a estimate of the significance of the related dictionary component, so weighing the deformation coefficients $\boldsymbol{\alpha}$ by $\mathbf{w}^{-1}$ induce a reduced penalty for the most relevant components. The non-rigid coefficients are estimated in a closed form solution as:

$$\boldsymbol{\alpha} = \left(\mathbf{Y}^T\mathbf{Y} + \lambda\cdot diag(\mathbf{w}^{-1})\right)^{-1}\mathbf{Y}^T\mathbf{X}\ . \tag{10}$$

5. Experimental results

The proposed 3DMM has been evaluated in two sets of experiments. First, we compared the 3DMM constructed from a dictionary of learned components (DL-3DMM), and its counterpart obtained using PCA decomposition (PCA-3DMM) (Sect. 5.1). Then, we compared the two solutions in the task of face recognition from images that show jointly large pose and expression variations (Sect. 5.2). Both the experiments have been performed on the BU-3DFE dataset.

5.1. Reprojection and reconstruction error

This experiment aims at comparing the proposed DL-3DMM and its PCA-3DMM counterpart. In so doing, we also aim at evaluating how the 3DMM changes when the different parameters involved in its construction are modified. To separate the fitting error from possible inaccuracies induced by landmarks detection and pose estimation, we used a projection of the 3D scans as ground truth. In particular, all the 3D scans (the whole meshes) are projected onto the 2D plane using a same reference projection matrix ($\mathbf{P}_{ref}$). For the sake of generality, the projection $\mathbf{P}_{ref}$

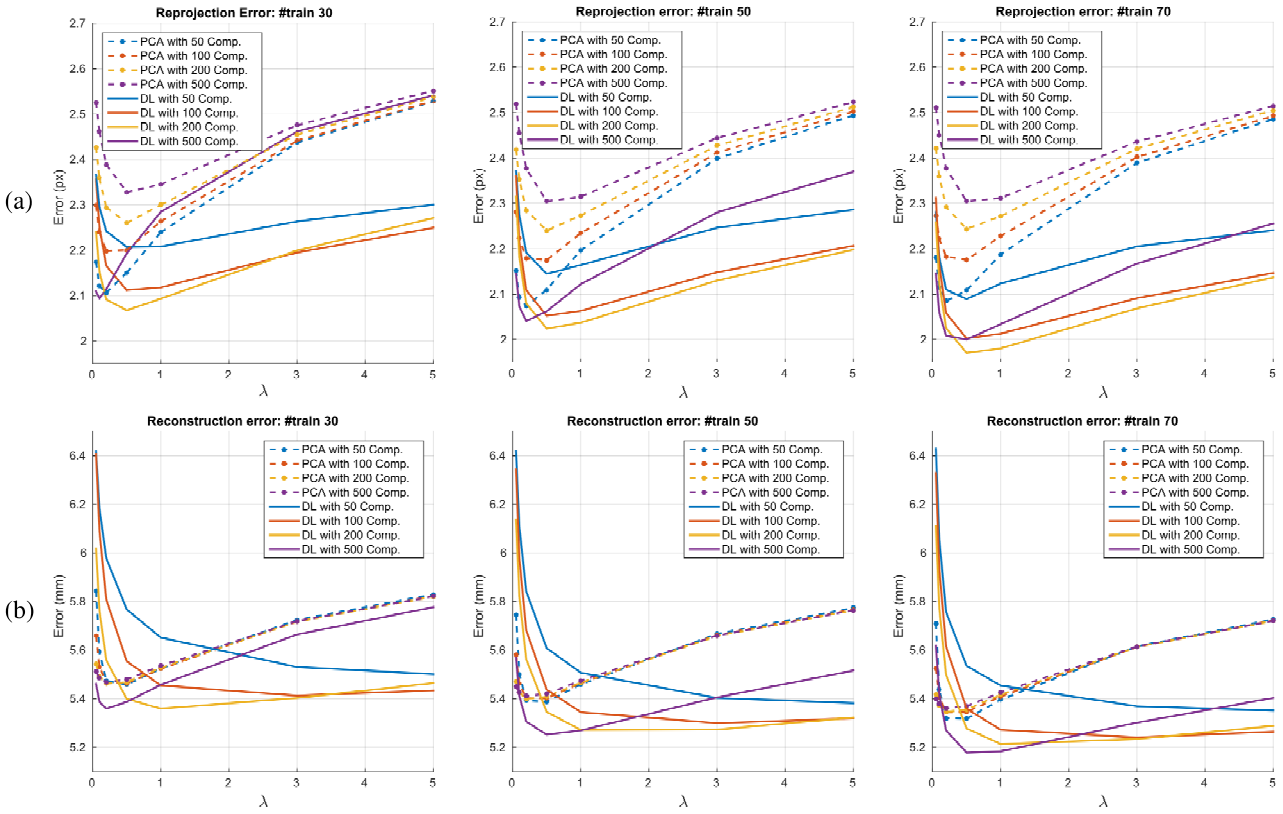


Figure 5. Error vs. regularization parameter λ for: (a) Reprojection; (b) Reconstruction. In both the cases, plots are reported for the 3DMM learned on training sets that include 30, 50 and 70 identities, respectively, from left to right. Each plot compares, for different number of components, the results obtained for the DL-3DMM (DL), and for its PCA-3DMM counterpart (PCA).

has been chosen in a intermediate pose between a frontal view and a side view. The 3DMM is then fit following the approach described in Sect. 4, using as 2D landmarks the points of the projected models selected by $\mathbf{I}_v$ (set $\mathbf{l}$ in Eq. (8)). The distance between the ground-truth model and the deformed one, projected on the 2D plane, is used to measure the 2D *reprojection* error. The same distance in 3D is instead used to measure the 3D *reconstruction* error. These two distances are computed overall the vertices of the mesh. The deformation of the 3DMM guided by the landmark correspondences should make the model as similar as possible to the 3D ground-truth scan. The Euclidean distance is used as error measure in both the cases, averaged over the number of vertices. All the above experiments have been performed by splitting the data into a train and a test fold, and then using cross-validation. In particular, we used all the 100 identities of the BU-3DFE and considered 30-70, 50-50 and 70-30 random partitions between the train and test folds. In this way, the identities used in the test are completely separated from the identities used in the train for the 3DMM construction. This process is repeated ten times and results are averaged across the ten trials.

Results for the reprojection and reconstruction errors are reported in Fig. 5(a) and (b), respectively. In both the cases, plots refer to the 3DMM learned on training sets with 30, 50 and 70 identities, respectively, from left to right. Each plot compares, for different number of components, the results obtained for the DL-3DMM constructed using coding on a learned dictionary (DL), and its PCA-based counterpart (PCA). The error (in $pixels$ for reprojection, in mm for reconstruction) is plotted versus the regularization parameter λ appearing in Eq. (9). Results show that an optimal value of λ can be found at about 0.5, and this value is almost stable across different sizes of the training set. For the DL-3DMM the error decreases up to 200 components. The error for the DL 3DMM is also smaller than the error observed for the PCA-3DMM.

5.2. Face recognition

The effectiveness of the proposed 3DMM construction and fitting approach has been also evaluated in face recognition from images with challenging poses and expressions. To this end, we included in the gallery the frontal neutral images as rendered from the 3D scans of the BU-3DFE,

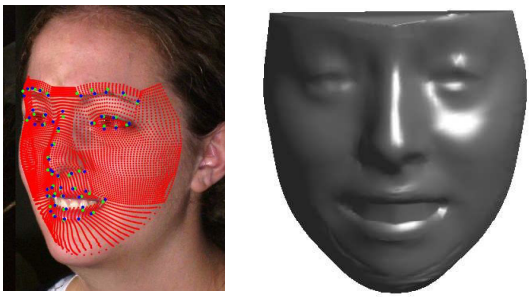


Figure 6. 3DMM fitting example: (a) Face image with the detected landmarks (in blue), their reprojected position (in green), and the projected mesh (red dots); (b) Fitted 3DMM.

	#train 30		#train 50		#train 70	
	DL	PCA	DL	PCA	DL	PCA
Neutral	55.0	**55.9**	**56.1**	55.7	57.9	57.9
Angry	**46.3**	45.5	**47.6**	46.9	**49.5**	49.2
Disgust	**47.6**	44.9	**48.9**	46.7	49.7	49.7
Fear	**51.4**	49.7	**51.7**	49.9	**53.5**	53.2
Happiness	**51.0**	49.3	**50.5**	49.9	**53.8**	53.0
Sadness	**45.1**	44.8	**45.3**	44.6	48.5	**50.5**
Surprise	**39.3**	37.1	**40.4**	38.5	**41.7**	40.9
All	**47.2**	45.5	**47.9**	46.4	**49.8**	49.5

Table 1. Face recognition accuracy for probes with left / right pose and for the different expressions. Results for the proposed DL-3DMM and its PCA based counterpart are reported.

while the probes are all the images where the subjects exhibit pose and expression variations (namely, *angry, disgust, fear, happy, sad, surprise*). We remark here these images also show a left / right side pose of the subjects, as illustrated in Fig. 1(b), so that the combination of pose and expression makes the recognition scenario very difficult. The landmark detector in [25] is run on gallery and probe images, and the 3DMM is fit on them. We then exploit the projection on the image of the 3DMM to sample RGB values in correspondence of the projected vertices, so as to render a frontal view. These latter images are then used to compute Local Binary Patterns (LBP) descriptors over patches centered on the projected vertices of the 3DMM instead of considering a regular grid over the image. Finally, a nearest-neighbor classifier is used to match the LBP descriptors in the probe and gallery images. Also for this experiment, the same cross-validation approach reported in Sect. 5.1 has been used.

A fitting example is reported in Fig. 6: In (a), the detected and reprojected landmarks are evidenced in blue and green, respectively, while the projected mesh is given with red dots; In (b), the deformed DL-3DMM is shown.

Table 1 reports the face recognition results obtained for probes with left / right pose and for the different expressions. Results for the proposed DL-3DMM and its PCA based counterpart are compared. In general, the accuracy increases with the size of the training set, with the DL-3DMM performing better than the PCA-3DMM in most of the cases. In particular, the gap between the two solutions is more marked in the case of small training sets.

Finally, face recognition has been also studied as a function of the expression intensity from *neutral* (level-0), to expressive (levels from 1 to 4), using different numbers of training subjects. Results for the DL-3DMM show a decrease of recognition passing from level-0 to level-2 of about 8% to 10%, while an almost stable behaviour is observed going from level-2 to level-3 and -4. Compared to the PCA, the DL-3DMM obtains better performance in the case of 30 and 50 training subjects, with a gap of about 2% to 3%, which increases with expression intensity. For the case of 70 training subjects, the difference between the two solutions is mostly irrelevant, apart for level-4 intensity where the DL-3DMM clearly outperforms PCA-3DMM.

6. Discussion and future work

In this work, we have proposed a dictionary learning method for constructing a 3DMM, and we have proved its usefulness for face recognition in the case of images that combine challenging pose and expression variations. Compared to traditional methods based on PCA, our solution has the advantage of permitting more localized variations of the 3DMM that can better adapt to expressive faces. Differently from existing 3DMM, we also propose to use a training set that includes a large spectrum of facial variations, in terms of ethnicity, age, and facial expressions. A new method for establishing dense correspondence between annotated facial scans in the training is also proposed, which permits us to process difficult expressions. The proposed DL-3DMM has then been cast to a rigid / non-rigid deformation framework, which includes pose estimation and regularized ridge-regression fitting, thus allowing its experimentation in reconstruction and recognition experiments. Results show the DL-3DMM improves the standard solution based on PCA decomposition in both the experiments.

As future work, we will explore the possibility to exploit the dictionary learning in a sparsity coding fashion. In addition, we aim to experiment the proposed 3DMM with more effective facial features, and on new face datasets in the wild. Finally, while in this work we rely on the landmarks provided with the BU-3DFE for establishing the dense correspondence in the training set, the existence of robust landmark detectors both in 2D [25] and 3D [20] will be exploited to derive a completely automatic solution.

Acknowledgment

This research was partially funded by IARPA's JANUS program (contract number 2014-14071600011).

References

[1] B. Amberg, R. Knothe, and T. Vetter. Expression invariant 3D face recognition with a morphable model. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–6, Amsterdam, The Netherlands, Sept. 2008. 1, 2, 3

[2] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Minneapolis, MN, USA, June 2007. 2, 3

[3] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. *Computer Graphics Forum*, 22(3):641–650, Nov. 2003. 1

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *ACM Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, Los Angeles, CA, USA, Aug. 1999. 1, 3, 5

[5] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, Sept. 2003. 1, 2

[6] T. Bolkart and S. Wuhrer. 3D faces in motion: Fully automatic registration and statistical analysis. *Computer Vision and Image Undrstanding*, 131(2):100–115, Feb. 2015. 4

[7] A. Brunton, T. Bolkart, and S. Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *European Conf. on Computer Vision (ECCV)*, pages 297–312, Zurich, Switzerland, Sept. 2014. 2

[8] 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 Undrstanding*, 128(11):1–17, Nov. 2014. 2

[9] J. Bustard and M. Nixon. 3D morphable model construction for robust ear and face recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2582–2589, San Francisco, CA, USA, June 2010. 2

[10] M. De Smet and L. J. Van Gool. Optimal regions for linear model-based 3D face reconstruction. In *Asian Conf. on Computer Vision (ACCV)*, pages 276–289, Queenstown, New Zealand, Nov. 2010. 4

[11] L. G. Farkas and I. R. Munro. *Anthropometric Facial Proportions in Medicine*. Thomas Books, Springfield, IL, USA, 1987. 1

[12] 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 Trans. on Pattern Analysis and Machine Intelligence*, 29(4):640–649, Apr. 2007. 2

[13] R. Kimmel and J. Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Science*, 95(15):8431–8435, July 1998. 4

[14] X. Lu and A. Jain. Deformation modeling for robust 3D face matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(8):1346–1357, Aug. 2008. 4

[15] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Int. Conf. on Machine Learning (ICML)*, pages 689–696, Montreal, Quebec, Canada, June 2009. 5

[16] I. Masi, C. Ferrari, A. Del Bimbo, and G. Medioni. Pose independent face recognition by localizing local binary patterns via deformation components. In *Int. Conf. on Pattern Recognition (ICPR)*, pages 4477–4482, Stockholm, Sweden, Aug. 2014. 1, 6

[17] T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, and C. Theobalt. Sparse localized deformation components. *ACM Trans. Graphics*, 32(6):179:1–179:10, Nov. 2013. 5

[18] A. Patel and W. Smith. 3D morphable face models revisited. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1327–1334, Miami Beach, FL, USA, June 2009. 2, 4

[19] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, pages 296–301, Genoa, Italy, Sept. 2009. 2, 3

[20] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris. 3D facial landmark detection under large yaw and expression variations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(7):1552–1564, 2013. 8

[21] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 986–993, San Diego, CA, USA, June 2005. 1, 2

[22] A. Salazar, S. Wuhrer, C. Shu, and F. Prieto. Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, 25(4):859–879, May 2014. 4

[23] D. Shahlaei and V. Blanz. Realistic inverse lighting from a single 2D image of a face, taken under unknown and complex lighting. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Ljubljana, Slovenia, May 2015. 1

[24] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, Mar. 1987. 1

[25] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, Portland, OR, USA, June 2013. 6, 8

[26] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3D facial expression database for facial behavior research. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 211–216, Southampton, UK, Apr. 2006. 3

[27] L. Zhang, Y. Wang, S. Wang, D. Samaras, S. Zhang, and P. Huang. Image-driven re-targeting and relighting of facial expressions. In *Computer Graphics Int. (CGI)*, pages 11–18, Stony Brook, NY, USA, June 2005. 1

[28] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3D morphable model fitting. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Ljubljana, Slovenia, May 2015. 1, 2