A Dictionary Learning-Based 3D Morphable Shape Model

Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, *Senior Member, IEEE*, and Alberto Del Bimbo

***Abstract*—Face analysis from 2D images and videos is a central task in many multimedia applications. Methods developed to this end perform either face recognition or facial expression recognition, and in both cases results are negatively influenced by variations in pose, illumination, and resolution of the face. Such variations have a lower impact on 3D face data, which has given the way to the idea of using a 3D morphable model as an intermediate tool to enhance face analysis on 2D data. In this paper, we propose a new approach for constructing a 3D morphable shape model (called DL-3DMM) and show our solution can reach the accuracy of deformation required in applications where fine details of the face are concerned. For constructing the model, we start from a set of 3D face scans with large variability in terms of ethnicity and expressions. Across these training scans, we compute a point-to-point dense alignment, which is accurate also in the presence of topological variations of the face. The DL-3DMM is constructed by learning a dictionary of basis components on the aligned scans. The model is then fitted to 2D target faces using an efficient regularized ridge-regression guided by 2D/3D facial landmark correspondences in order to generate pose-normalized face images. Comparison between the DL-3DMM and the standard PCA-based 3DMM demonstrates that in general a lower reconstruction error can be obtained with our solution. Application to action unit detection and emotion recognition from 2D images and videos shows competitive results with state of the art methods on two benchmark datasets.**

***Index Terms*—Action unit detection, dictionary learning, dense correspondence, emotion recognition, 3D morphable model.**

I. Introduction

In recent years, the analysis of human faces has become increasingly relevant, with a variety of potential computer vision and multimedia applications. Examples include human identification based on face [1]–[3], emotional state detection [4], [5], enhanced human-computer interaction using facial pose and expression [6]–[10], facial expression detection for medical assistance or investigation [11], [12], prediction of drivers cognitive load [13], [14], just to cite some of the most studied.

Face recognition and facial expression recognition are central in most of these applications. In a broad sense, face recognition performs *coarse* grained (*inter-class*) face analysis, where face variations that separate different identities are accounted for. Convincing results have been obtained for 2D still images and videos, where 2D large and heterogeneous corpora can be easily collected, with the ultimate effect of making machine learning tools work effectively [15]–[17]. Conversely, in applications that recognize facial expressions or Action Units (AU), *fine* grained (*intra-class*) face analysis is required, where subtle and local variations of the face occur under the action of groups or individual facial muscles. In this case, 2D data manifest more evident limitations, and performing face analysis in 3D can be convenient [18], [19]. However, acquiring high-quality 3D data is more expensive and difficult than for 2D. For these reasons, using 2D/3D solutions becomes a suitable alternative. Even a limited amount of 3D data can indeed be exploited to produce an enhanced 2D representation.

In this framework, a potentially interesting idea is that of learning a generic 3D face model capable of generating new face instances with plausible shape and appearance. This can be done by capturing the face variability in a training set of 3D scans and constructing a statistical face model that includes an average component and a set of learned principal components of deformation. Such a model would allow either to generate new face instances, or deform and fit to 2D or 3D target faces.

Blanz and Vetter [20] first proposed to create a 3D morphable model (3DMM) from a set of exemplar 3D faces and showed its potential and versatility. 3DMM and its variants have been used with some success in coarse grained recognition applications, such as pose robust face recognition [21], [22] and 3D face recognition [23]. However, in the literature there are no convincing examples of 3DMMs applied to fine grained face analysis. This is mainly related to the difficulty existing 3DMMs have with coping with noise, local deformations and topology variations of the face. The ultimate motivation for this can be found in the two elements that are at the base of constructing a valid 3DMM: the data used for training, and the statistical tools that are applied to the data.

Intuitively, the capability of generating new face instances with realistic traits mainly depends on the variance in the training data. A training dataset that includes a significant sample of

Manuscript received October 20, 2016; revised February 21, 2017 and April 19, 2017; accepted May 8, 2017. Date of publication May 23, 2017; date of current version November 15, 2017. This work was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Cha Zhang. *(Corresponding author: Claudio Ferrari.)*

The authors are with the Department of Information Engineering, University of Firenze, Firenze 50139, Italy (e-mail: claudio.ferrari@unifi.it; giuseppe.lisanti@unifi.it; stefano.berretti@unifi.it; alberto.delbimbo@unifi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2707341

human face variability in terms of gender, ethnicity, age, and facial expressions is a fundamental prerequisite. However, a point-to-point dense correspondence is required to preserve the anatomical meaning of every point across all the training scans. This is a difficult task especially for expressive faces; in most of the cases, the use of scans with large differences has not been exploited to its full potential. On the other hand, PCA has been the tool applied for capturing the statistical variability in the training data in most state of the art works, with few other solutions explored. While this is possible only for scans in dense correspondence, if the correspondence is not accurate, most principal components will include noise.

In this paper, we propose a new approach to the construction of a 3D Morphable Shape Model (note that, even though we consider the sole shape component, throughout the paper we will use the term 3DMM for our solution). The proposed model is capable of capturing much of the large variability of human faces, thus opening the way to its use in fine grained face analysis. It is grounded in three distinct contributions:

1) a new method to establish a dense correspondence between scans even in the case of expressions that include topological variations such as open/closed mouth;
2) a new approach to capturing the statistical variability in training data that, instead of exploiting standard PCA, learns a dictionary of deformations from the deviations between each 3D scan and the average model computed on the vertex positions of the densely aligned training scans. We refer to this new model composed by the average model component and the learned basis of deviations as DL-3DMM;
t3) the application of 3DMM to the fine grained tasks of AU detection and emotion recognition from 2D images. To this end, we also propose an efficient fitting approach that only relies on the correspondence between 2D and 3D landmarks of the face, and avoids a costly iterative optimization by estimating the model parameters through a closed form solution.

In the experiments, we demonstrate the DL-3DMM compares favorably with respect to the standard PCA-based 3DMM in terms of reconstruction error. In addition, results show the effective applicability of 3DMM to facial AU detection and emotion recognition, with comparable performance to state of the art 2D methods, even using off-the-shelf image descriptors and learning solutions.

Some preliminary ideas of this work appeared in [24]. With respect to [24], here we extend the dense alignment approach to the whole face, and to the case where facial landmarks used for face partitioning are automatically detected. In addition, we propose a new experiment that includes a cross-dataset comparison and finally we demonstrate the applicability of our DL-3DMM to AU detection and emotion recognition.

The rest of the paper is organized as follows: in Section II, we review the state of the art on 3DMM construction and its application in image face analysis; in Section III, we present the method for determining dense correspondence between the 3D scans of a training set with a large spectrum of face variations; the DL-3DMM construction using dictionary learning is proposed in Section IV; in Section V, we present the 3DMM fitting method; Section VI introduces to the application of the 3DMM to AU detection and emotion recognition; in Section VII, we compare the DL-3DMM to the PCA-3DMM, and present their results for AU detection and emotion recognition; finally, discussion and conclusions are reported in Section VIII.

II. Related Work

In their seminal work, Blanz and Vetter [20] first presented a complete solution to derive a 3DMM by transforming the shape and texture from a training set of 3D face scans into a vector space representation based on PCA. A gradient-based optical flow algorithm was used to establish dense correspondence between pairs of 3D scans taking into account for texture and shape values simultaneously. A reference scan was then used to transfer correspondences across scans. However, the training dataset had limited face variability (200 neutral scans of young Caucasian individuals were included), thus reducing the capability of the model to generalize to different ethnicity and non-neutral expressions. Despite these limitations, the 3DMM has proved its effectiveness in image face analysis, also inspiring most of the subsequent work, with applications to computer graphics for face inverse lighting [25], [26] and reanimation [27], craniofacial surgery [28], 3D shape estimation from 2D image face data [29], 3D face recognition [23], pose robust face recognition [21], [22], etc.

The 3DMM was further refined into the Basel Face Model by Paysan *et al.* [30]. This offered higher shape and texture accuracy thanks to a better scanning device, and a lower number of correspondence artifacts using an improved registration algorithm based on the non-rigid iterative closest point (ICP) [31]. However, since non-rigid ICP cannot handle large missing regions and topological variations, expressions were not accounted for in the training data also in this case. In addition, both the optical flow used in [20] and the non-rigid ICP method used in [23], [30] were 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 the detected correspondences, and ultimately the final 3DMM. The work by Booth *et al.* [15], introduced a pipeline for 3DMM construction. Initially, dense correspondence was estimated applying the non-rigid ICP to a template model. Then, the so called LSFM-3DMM was constructed using PCA to derive the deformation basis on a dataset of 9,663 scans with a wide variety of age, gender, and ethnicity. Though the LSFM-3DMM was built from the largest dataset compared to the current state-of-the-art, the face shapes still were in neutral expression.

Following a different approach, Patel and Smith [32] showed that Thin-Plate Splines (TPS) and Procrustes analysis can be used to construct a 3DMM. Procrustes analysis was used to establish correspondence between a set of 104 manually labelled landmarks of the face, and the mean coordinates of these landmarks were used as anchor points. A complete deformable model was then constructed by warping the landmarks of each sample to the anchor points and interpolating the regions between landmarks using TPS. Finally, consistent resampling was performed across all faces, but using the estimated surface between landmarks rather than the real one. In [33], Cosker *et al.* described a framework for building a dynamic 3DMM, which extended static 3DMM construction by incorporating dynamic data. This was obtained by proposing an approach based on Active Appearance Model and TPS for non-rigid 3D mesh registration and correspondence. Results showed this method overcomes optical flow based solutions that are prone to temporal drift. Brunton *et al.* [34], instead, proposed a statistical model for 3D human faces in varying expression. The approach decomposed the face using a wavelet transform, and learned many localized, decorrelated multilinear models on the resulting coefficients. In [35], Lüthi *et al.* presented a Gaussian Process Morphable Model (GPMM), which generalizes PCA-based Statistical Shape Models (SSM). GPMM was defined by a Gaussian process, which makes it inherently continuous. Further, it can be specified using arbitrary positive definite kernels, which makes it possible to build shape priors, even in the case where many examples to learn an SSM are not available.

3DMM has been used at coarse level for face recognition and synthesis. In one of the first examples, Blanz and Vetter [21] used their 3DMM to simulate the process of image formation in 3D space, and estimated 3D shape and texture of faces from single images for face recognition. Later, Romdhani and Vetter [36] used the 3DMM for face recognition by enhancing the deformation algorithm with the inclusion of various image features. In [37], Yi *et al.* used the 3DMM to estimate the pose of a face image with a fast fitting algorithm. This idea was extended further by Zhu *et al.* [38], who proposed fitting a dense 3DMM to an image via Convolutional Neural Network. Grupp *et al.* [39] fitted the 3DMM based exclusively on facial landmarks, corrected the pose of the face and transformed it back to a frontal 2D representation for face recognition. Hu *et al.* [40] proposed a Unified-3DMM that captures intra personal variations due to illumination and occlusions, and showed its performance in 3D-assisted 2D face recognition for scenarios where the input image is subjected to degradations or exhibits intra-personal variations.

In all these cases, the 3DMM was used mainly to compensate for the pose of the face, with some examples that performed also illumination normalization. Expressions were typically not considered. Indeed, the difficulty in making 3DMM work properly in fine face analysis applications is confirmed by the almost complete absence of methods that use 3DMM for expression recognition. Among the few examples, Ramanathan *et al.* [41] constructed a 3D Morphable Expression Model incorporating emotion-dependent face variations in terms of morphing parameters that were used for recognizing four emotions. Ujir and Spann [42] combined the 3DMM with Modular PCA and Facial Animation Parameters (FAP) for facial expression recognition, but the model deformation was due more to the action of FAP than to the learned components. In [43], Cosker *et al.* used a dynamic 3DMM [44] to explore the effect of linear and non-linear facial movement on expression recognition through a test where users evaluated animated frames. Huber *et al.* [45] proposed a cascaded-regressor based face tracking and a 3DMM shape fitting for fully automatic real-time semi dense 3D face reconstruction from monocular in-the-wild videos.

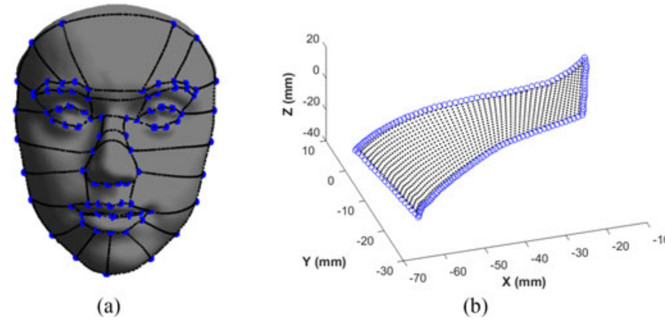


Fig. 1. (a) Face scan of the BU-3DFE with the 87 landmarks (in blue) and the geodesic paths used to connect some of them (in black). These paths partition the face into a set of non-overlapping regions. (b) Geodesic contour of the *cheek/zygoma* region on the right side of the face. The geodesic contour is resampled so that points on it (circles in the plot) are at the same geodesic distance from each other. The interior of the region is also resampled using linear paths on the surface (dots in the plot), which connect corresponding points on opposite sides of the contour.

III. Finding 3D Dense Correspondence

Given a training set, finding a dense point-to-point correspondence between the vertices of 3D scans can be seen as a sort of mesh re-parametrization where corresponding points must have the same anatomical reference. The limited number of facial points detectable with sufficient accuracy, and the presence of large regions with strong photometric variations, self-occlusions, facial expressions and changes in the topology of the face surface (as in the case of mouth-closed / mouth-open), make this problem highly complex.

In our approach, similarly to Patel and Smith [32], we initially rely on a set of landmarks to establish a correspondence between salient points of the face [see Fig. 1(a)]. However, differently from [32], where warping and TPS interpolation is applied between the average landmarks, we interpolate and sample the scan surface, region-by-region, while maintaining a dense correspondence. We first partition the face into a set of regions using geodesic paths between facial landmarks, applying the variant of the Fast Marching algorithm on triangular mesh manifolds of [46], and resample the geodesics with a predefined number of points posed at equal geodesic distance. As an example, Fig. 1(b) shows (with circles), the sampled points of the geodesic contour delimiting the cheek/zygoma region comprised between the nose and the face boundary on the right. Hence, we sample the surface of the face regions 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 linear path on the surface [47]. This line is then sampled at the desired resolution, as illustrated in Fig. 1(b). Being based on the annotated landmarks and their connections, this approach proved to be robust to facial expressions. In particular, the presence of landmarks which delimit the internal and external border of the lips, makes it possible to maintain such region correspondence also across faces with mouth-closed/mouth-open expressions. While the method of [32] is only able to estimate the real surface, in our case, we are able to interpolate and sample the true surface of the face scans, region-by-region, maintaining a dense correspondence and do not require an average model as in [32]. With respect to the solutions in [20], [30], [31] our approach does not require a reference face model, that could request a new face parametrization. It only requires that training faces are labeled with a set of landmarks, that is easily obtained with good accuracy using available detectors both in 2D [48], [49] and 3D [50].

Learning a 3DMM requires a training set of 3D face scans with high variability in terms of gender, age and ethnicity. Since we aim to generalize to expressive data, including scans with facial expressions is also important. To this end, we used the publicly available Binghamton University 3D Facial Expression dataset (BU-3DFE) [51] as training set. This dataset includes a balanced sample of human face variability and facial expressions and has been largely employed for 3D expression/face recognition. In particular, the BU-3DFE contains scans of 44 females and 56 males, with age ranging from 18 to 70 years old, acquired in a neutral plus six different expressions: anger, disgust, fear, happiness, sadness, and surprise. Apart from neutral, all the other facial expressions were acquired at four levels of intensity, from low to exaggerated (2500 scans in total). The subjects are distributed across different ethnic groups or racial ancestries, including *White*, *Black*, *Indian*, *East-Asian*, *Middle East Asian*, and *Hispanic-Latino*. The 83 facial landmarks annotated and released with the BU-3DFE provide correspondence across the training faces for a limited set of anchor points in correspondence to the distinguishing traits of the face.

Four additional landmarks located in the forehead have been derived from this initial set using anthropometric considerations on face proportions [52]. The overall set of 87 landmarks is shown with blue spheres on the face scan in Fig. 1(a). It is evident that these landmarks delimit salient parts of the face: the eyebrows, the eyes, the upper and lower lips, the nose, and the face boundary. By connecting selected pairs of landmarks through geodesic paths, we identified 12 regions in each side of the face (comprising the super-orbitali, eyebrow, eye, cheek, jaw and chin), plus 9 regions covering the middle part of the face (including the lips, the region between the upper lip and the nose, the nose, the region between the eyes, and the forehead). As a result, each face was partitioned into 33 regions, each delimited by a closed geodesic contour passing through a set of landmarks, as shown in Fig. 1(a).

IV. DL-3DMM Construction

Once a dense correspondence is established across the training data, we build our DL-3DMM by learning a dictionary of deformation components exploiting the *Online Dictionary Learning for Sparse Coding* technique [53]. Learning is performed in an unsupervised way, without exploiting any knowledge about the data (e.g., identity or expression labels).

Let N be the set of training scans, as obtained in Section III, each with m vertices. Each scan is represented as a column vector $\mathbf{f}_i \in \mathbb{R}^{3m}$, whose elements are the linearized X, Y, Z coordinates of all the vertices, that is

$$\mathbf{f}_i = [\, X_{i,1}\, Y_{i,1}\, Z_{i,1}\, \ldots\, X_{i,m}\, Y_{i,m}\, Z_{i,m}\,]^T \in \mathbb{R}^{3m} \;.$$

The average model $\mathbf{m}$ of the training scans is computed as

$$\mathbf{m} = \frac{1}{|N|}\sum_{i=1}^{|N|} \mathbf{f}_i \;. \tag{1}$$

Then, for each training scan $\mathbf{f}_i$, we compute the field of deviations $\mathbf{v}_i$ with respect to the average model $\mathbf{m}$

$$\mathbf{v}_i \leftarrow \mathbf{f}_i - \mathbf{m}\;,\;\; \forall\, \mathbf{f}_i \in N\;. \tag{2}$$

In the classic 3DMM framework [20], new 3D shapes are generated by deforming the average model $\mathbf{m}$ with a linear combination of the principal components. In this work, instead, we propose to learn a set of deformation components through dictionary learning. In particular, the dictionary atoms are learnt from the field of deviations $\mathbf{v}_i$. Then, we morph the average model exploiting a linear combination of the dictionary atoms. Note that the PCA model is also constructed on the training set $\mathbf{v}_i$.

Dictionary learning is usually cast as an ℓ_1-regularized least squares problem [53]. However, since the learnt directions are used to deform the average model, the sparsity induced by the ℓ_1 penalty can lead to a noisy or, in the worst case, a discontinuous or punctured model. We thus decided to formulate the dictionary learning as an *Elastic-Net* regression. 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, while forcing smoothness. By defining $\ell_{1,2}(\mathbf{w}_i) = \lambda_1 \|\mathbf{w}_i\|_1 + \lambda_2 \|\mathbf{w}_i\|_2$, where λ_1 and λ_2 are, respectively, the sparsity and regularization parameters, we can formulate the problem as

$$\min_{\mathbf{w}_i,\,\mathbf{D}} \frac{1}{|N|}\sum_{i=1}^{|N|}\Big(\|\mathbf{v}_i - \mathbf{D}\mathbf{w}_i\|_2^2 + \ell_{1,2}(\mathbf{w}_i)\Big) \tag{3}$$

where the columns of the dictionary $\mathbf{D} \in \mathbb{R}^{3m\times k}$ are the basis components, $\mathbf{w}_i \in \mathbb{R}^k$ are the coefficients of the dictionary learning, and k is the number of basis components of the dictionary. The number of components (dictionary atoms) must be defined a priori. Instead, the set of coefficients $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k] \in \mathbb{R}^{k\times k}$ is obtained as the cumulated sum of the coefficients at each iteration of the dictionary learning. The coefficients of the matrix $\mathbf{W}$ are in general concentrated on the diagonal [53], and represent the contribution of the k-th basis element in reconstructing the training vectors.

The above minimization can be rewritten as a joint optimization problem with respect to the dictionary $\mathbf{D}$ and the coefficients $\mathbf{W}$, and solved by alternating between the two variables, minimizing over one while keeping the other one fixed [53]. The average model $\mathbf{m}$, the dictionary $\mathbf{D}$ and the diagonal elements of the matrix $\mathbf{W}$, namely the vector $\hat{\mathbf{w}} \in \mathbb{R}^k$, constitute our *Dictionary Learning based* 3DMM (DL-3DMM).

V. Efficiently Fitting the DL-3DMM

Fitting a 3DMM to a 2D face image allows a coarse 3D reconstruction of the face. To this end, estimating the 3D pose of the face, and the correspondence between 3D and 2D landmarks are prerequisites. In the following, both the average model and the basis components of the learned dictionary will be represented in $\mathbb{R}^{3\times m}$, rather than in $\mathbb{R}^{3m}$, and we refer to them as $\hat{\mathbf{m}}$ and $\hat{\mathbf{D}}$, respectively.

In order to estimate the pose, we detect a set of 49 facial landmarks $\mathbf{l} \in \mathbb{R}^{2\times 49}$ on the 2D face image using the technique proposed in [48] (see Fig. 2). An equivalent set of vertices $\mathbf{L} = \hat{\mathbf{m}}(\mathbf{I}_v) \in \mathbb{R}^{3\times 49}$ is manually annotated on the average 3D model, where $\mathbf{I}_v$ is the set of indices of the vertices corresponding to the landmark locations. Under an affine camera model [22], the relation between $\mathbf{L}$ and $\mathbf{l}$ is

$$\mathbf{l} = \mathbf{A} \cdot \mathbf{L} + \mathbf{T} \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{2\times 3}$ contains the affine camera parameters, and $\mathbf{T} \in \mathbb{R}^{2\times 49}$ is the translation on the image. To recover these parameters, firstly, we subtract the mean from each set of points and recover the affine matrix $\mathbf{A}$ solving the following least squares problem:

$$\arg\min_{\mathbf{A}} \|\mathbf{l} - \mathbf{A} \cdot \mathbf{L}\|_2^2 \tag{5}$$

for which the solution is given by $\mathbf{A} = \mathbf{l} \cdot \mathbf{L}^{+}$, where $\mathbf{L}^{+}$ is the pseudo-inverse matrix of $\mathbf{L}$. We can estimate the affine matrix with a direct least squares solution since, by construction, facial landmark detectors assume a consistent structure of the 3D face parts so they do not permit outliers or unreasonable arrangement of the face parts (e.g., nose landmarks cannot stay above the eyes). Finally, the 2D translation can be estimated as $\mathbf{T} = \mathbf{l} - \mathbf{A} \cdot \mathbf{L}$. Thus, the estimated pose $\mathbf{P}$ is represented as $[\mathbf{A}, \mathbf{T}]$ and used to map each vertex of the 3DMM onto the image.

Using the learned dictionary $\hat{\mathbf{D}} = [\hat{\mathbf{d}}_1, \ldots, \hat{\mathbf{d}}_k]$, we find the coding that non-rigidly transforms the average model $\hat{\mathbf{m}}$ such that the projection minimizes the error in correspondence to the landmarks. The coding is formulated as the solution of a regularized *Ridge-Regression* problem

$$\arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v) - \sum_{i=1}^{k} \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)\alpha_i \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \circ \hat{\mathbf{w}}^{-1} \right\|_2 \tag{6}$$

where $\circ$ is the Hadamard product. Since the pose $\mathbf{P}$, the basis components $\hat{\mathbf{d}}_i$, the landmarks $\mathbf{l}$, and $\hat{\mathbf{m}}(\mathbf{I}_v)$ are known, we can define $\hat{\mathbf{X}} = \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v)$ and $\hat{\mathbf{y}}_i = \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)$. By considering their linearized versions[1] $\mathbf{X} \in \mathbb{R}^{98}$ and $\mathbf{y}_i \in \mathbb{R}^{98}$ with $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]$, we can finally estimate the non-rigid coefficients which minimize the cost of (6), in closed form as follows:

$$\boldsymbol{\alpha} = \left(\mathbf{Y}^T\mathbf{Y} + \lambda \cdot \text{diag}(\hat{\mathbf{w}}^{-1})\right)^{-1} \mathbf{Y}^T\mathbf{X} \tag{7}$$

where $\text{diag}(\hat{\mathbf{w}}^{-1})$ denotes the diagonal matrix with vector $\hat{\mathbf{w}}^{-1}$ on its diagonal. The term $\hat{\mathbf{w}}^{-1}$ is used to associate a reduced cost to the deformation induced by the most relevant components. Indeed, weighting the deformation parameters $\boldsymbol{\alpha}$ with the inverse of the coefficients $\hat{\mathbf{w}}$, reduces the cost of the deformation induced by components $\hat{\mathbf{d}}_i$ with a large coefficient $\hat{\mathbf{w}}_i$, while the contribution of unstable and noisy components is bounded. In the classic PCA model, the same principle applies, but in this case the deformation components $\hat{\mathbf{d}}_i$ are represented by the PC, while the vector $\hat{\mathbf{w}}$ corresponds to the eigenvalues associated to the PC.

Note that the pose estimation and fitting steps are alternated; we experimentally found that cleaner reconstructions can be obtained by repeating the process while keeping a high λ. This is motivated by the fact that the initial 3D and 2D landmark layouts are likely to be very different due to the presence of expressions, and the pose can be coarsely estimated. In this scenario, the non-rigid deformation which fits the landmark locations is likely to excessively deform the model in the attempt of compensating also the error introduced by the pose. On the contrary, a high λ avoids to some extent this behavior and permits refinement of both the pose and the non-rigid deformation in the next step. Thus, a balance is required between the number of steps and the value of λ. We empirically found that the best configuration is repeating the process 2 times, with λ ranging from 0.0001 to 0.05. More than 2 repetitions do not produce appreciable improvement in the fitting.

A fitting example obtained using this solution is shown in Fig. 2. As a result, the 3D model is deformed according to the target face image, and the pose of the model can be normalized to obtain a frontalized face image. The vertices of the model can also be projected onto the rendered face; we can therefore compute image feature descriptors in repeatable positions across different faces exploiting the projected model vertices. The ultimate result of this procedure is an improved alignment of the image descriptors, which has been proved relevant in several face analysis applications [1], [54].

We exploit the technique presented in [55] to render a canonical frontal view of the face; the knowledge of the 3D face shape allows us to compute a pixel-wise inverse transformation, which associates to each pixel a 3D location in the coordinate system of the 3D model. Practically, once the 3D model is fit and projected onto the image, for each 3D vertex $v_j = (X_j, Y_j, Z_j)$, we know the coordinates (x_j, y_j) of the pixel corresponding to the projection of the vertex on the 2D image plane. Conversely, many pixels of the image have not a direct map in 3D, since they do not correspond to the projection of any 3D vertex. The 3D locations of these pixels can be estimated by fitting a function $h(x, y)$ across all the scattered pixels for which the 3D to 2D mapping is known. Then, defining Ω as the convex hull of the projected 3DMM, the 3D position $g_{u,v}$ of each pixel $(u, v) \in \Omega$ is estimated as

$$g_{u,v} = h(u, v)\ , \quad \forall\, (u, v) \in \Omega\ . \tag{8}$$

This back-projection from the 2D image plane to the 3D space permits us to associate the color of any pixel to a point estimated on the 3D model. The resulting rendered image is artifact free and also pixel-wise aligned across all the images since the transformation is computed pixel-by-pixel (see Fig. 2).

[1]The dimension 98 results from the concatenation of the coordinates of the 49 landmarks.

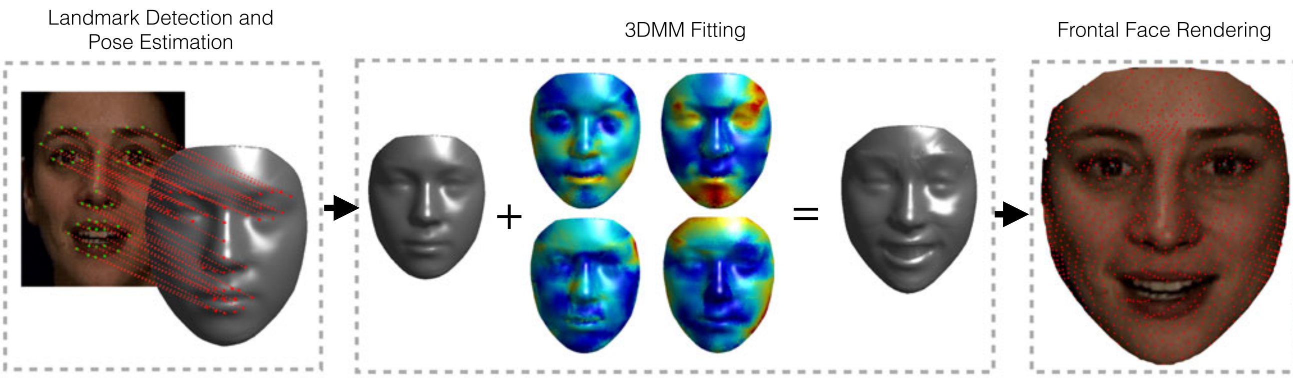


Fig. 2. Proposed 3DMM fitting and frontal face rendering: (left) the 3D head pose is estimated from the correspondence of 2D and 3D landmarks; (middle) the average 3D model is deformed using the basis components; (right) then, a frontal face image is rendered. A subsampling of the mesh vertices back projected onto the frontalized image is also shown. Note that, as a result of the fitting, no vertices fall inside the open mouth region.

Fig. 3. Example of fitting an expressive face with a 3DMM. The importance of including expressive scans in the training set can be appreciated: a 3DMM built without expressive scans fails in fitting the expressive face.

Moreover, the densely textured 3D model can also be used to render a face image in an arbitrary pose.

VI. 3DMM-Based Facial Analysis Application: AU Detection and Emotion Recognition

To the best of our knowledge, 3DMMs have not been used for the analysis of facial expressions; this can be reasonably ascribed to the difficulty of including expressive scans in the training data, which limits the capability of deforming a 3D model accurately in the presence of facial expressions, as shown in Fig. 3.

Facial expression analysis can be conducted mainly at two different levels: a finer one, i.e., *Action Unit (AU) detection*, which aims at detecting subtle movements of small parts of the face; and a more holistic one, which tries to classify the emotional state of the subject based on the whole face appearance, i.e., *Emotion recognition*.

Facial AUs are defined by the Facial Action Coding System (FACS) [56], which categorizes human facial movements based on the face appearance changes induced by the activity of the underlying muscles. The activation of an AU can thus be inferred from the observation of a face image. The AU detection task consists in deciding whether a particular AU is active or not in a given face image. Using this definition, in the literature, facial expressions have been systematically defined as the simultaneous activation of different AUs [57]. Facial expressions share common characteristics in the resulting face appearance and are also related to the emotional state of the subject showing the expression. Despite the precise definition, it is common that experts manually label face images referring to a set of standard discrete emotions, e.g., anger, fear, disgust, joy, happiness, relief, contempt, sadness and surprise.

To perform AU detection and Emotion Recognition a pipeline has been defined, which includes the following steps: 1) Image alignment; 2) Feature extraction; 3) Classifier training; 4) AU detection/Emotion recognition. The image alignment is performed by fitting the 3DMM to the face image so as to render a frontalized version using the method presented in Section V and summarized in Fig. 2. The face images are then described using LBP features [58], that are concatenated and projected to a lower dimensional space. Finally, classification/detection is performed using linear SVM classifiers, trained separately for each AU or emotion. The choice of using baseline image descriptors (LBP) and classifiers (LinearSVM) is motivated by the fact that our final goal is to assess the improvement that can be obtained using the 3DMM to enhance image description.

VII. Experimental Results

The proposed DL-3DMM has been evaluated in three sets of experiments. First, we investigate the modeling ability of the DL-3DMM compared with its PCA-based counterpart in terms of 3D to 2D fitting, and direct 3D to 3D fitting on the BU-3DFE. Then, we evaluate a cross-dataset fitting between the BU-3DFE and the Face Recognition Grand Challenge (FRGC v2.0) [59] dataset, by training on one dataset and testing on the other one, and vice versa. In both these experiments, two reference projection matrices are defined: $\mathbf{P}^f_{\text{ref}}$ simulates a subject facing the camera (*front view*); $\mathbf{P}^s_{\text{ref}}$ has been taken simulating a pose with approximately 45° in yaw (*side view*). The 3DMM is fit following the approach of Section V. For the direct 3D fitting, instead, we remove the projection $\mathbf{P}$ from (6) so as to perform the fitting directly in the original 3D space. Finally, we evaluate the 3DMM in the tasks of AU detection and emotion recognition, comparing it to baseline feature extraction approaches and state of the art solutions.

A. 3D Shape Reconstruction

We comparatively evaluate how the DL-3DMM and PCA-3DMM fit to a set of test images. Experiments were performed on the BU-3DFE, processed as illustrated in Section III so that scans are densely aligned with the same number of vertices. To train and test the 3DMMs, we split the scans into two halves based on subject identity (so that train and test identities are completely separated): one half of the scans is used to construct the average model $\mathbf{\hat{m}}$, the deformation components $\hat{\mathbf{d}}_i$, and the weights $\hat{\mathbf{w}}$ for both the DL-3DMM and the PCA-3DMM; the other half is used for test. This process is repeated 10 times on each train/test partition, and results are averaged across the trials.

To perform the 3D to 2D fitting, for each test scan we select the set of landmarks through the indices $\mathbf{I}_v$ and project them onto the 2D plane. These landmarks are used as a surrogate for the landmarks detected on a face image and allow both avoiding inaccuracies induced by detection and a misleading source of error not directly ascribable to the fitting.

Since the 2D landmarks are generated from the 3D scans, the original 3D data can be used as ground-truth of the model resulting from the fitting process. Based on this, we computed the 3D *reconstruction* error by accumulating the vertex-to-vertex Euclidean distance between the ground-truth scan and the deformed 3DMM. This measure exploits the knowledge of the exact correspondence between all the vertices of the 3D scans given by the dense alignment. Thus, the errors can be calculated by considering the distance between vertices with the same index in the meshes, without requiring any nearest vertex search. This is important, since in the presence of strong topological changes as determined by expressive scans, finding meaningful corresponding points for computing the errors is a complex task.

Reconstruction errors for three fitting conditions, namely, 3D-2D *front view*, 3D-2D *side view*, and 3D-3D are reported in Fig. 4(a), 4(b) and 4(c), respectively. The plots in the first row of the Figure compare the results obtained with the DL-3DMM and the PCA-3DMM as a function of the regularization parameter λ of (6) and for different number of components. The bar graph in the middle row shows the effect of varying the regularization parameter λ when the number of components is fixed at its best performing number, while in the bottom row it is shown the opposite, i.e., the effect of varying the number of components at the best regularization value. Results show that our DL-3DMM performs generally better than the PCA-3DMM. In particular, the two methods show a quite different behavior regarding the number of components used. For PCA-3DMM, we observe that increasing the number of components degrades the performance. This fact can be explained considering that 3D scans are noisy regardless of acquisition accuracy, and the alignment process can mitigate such nuisances only to some extent. Furthermore, it is likely that some PCs reflect less significant characteristics of the data. These facts eventually cause a drop of fitting accuracy due to the introduction of noisy and ineffective components, although regularized by their eigenvalues. This behavior is consistent with the concept of *compactness* of a model (i.e., the ability of explaining the most and significant variability of the data with the fewest number of components). On the opposite, the DL-3DMM improves its modeling ability with a larger number of components. This behavior is related to the fact that larger dictionaries allow more combinations of the atoms thus covering a wider range of possible deformations.

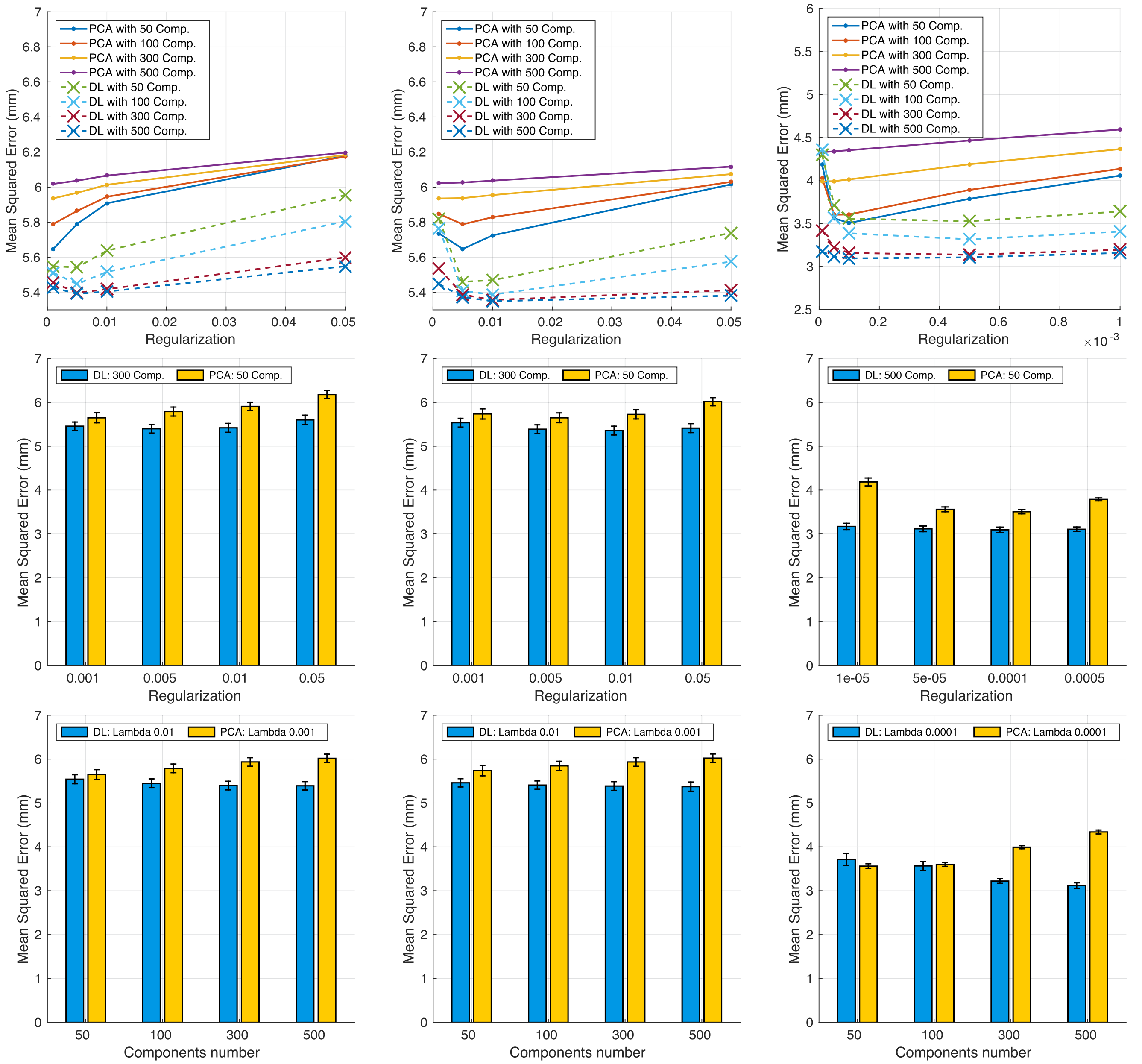
Results show that an optimal value of λ is about 0.01 and 0.001 for the DL and PCA methods, respectively. We point out here that despite producing the minimum error, using low regularization values to fit the 3DMM can occasionally result in noisy models; it is desirable instead to generate a model which is as smooth as possible. It can be observed from Fig. 4 that the reconstruction error is more stable across increasing λ values for the DL-3DMM rather than for the PCA-3DMM. It is then possible to choose a larger regularization value to ensure a smooth model, without renouncing modeling precision. This behavior is accentuated for increasing number of DL components.

Apart from the increased accuracy, since the fitting is quickly performed in closed form, we also note that the computational time still is acceptable even for a large number of components. We experimentally found that 2 repetitions of the whole fitting process of Section V take 17, 31, 103 and 185 ms for, respectively, 50, 100, 300, 500 components for both DL- and PCA-based 3DMM. We also found that after model deformation, the pose estimate is improved of about 0.5 degrees, with a final mean error of 5.0, 2.4, 4.1 degrees, respectively, for pitch, yaw and roll angles.

In Fig. 5 we show some examples of the deformation obtained using single dictionary atoms. Observe that DL components result in localized variations of the model, with a remarkable gap between different face parts. Moreover, by varying the magnitude of the deformation applied to the average model it is possible to generate new meaningful models. In Fig. 6 some examples of the 3DMM fitting, obtained using all the components, are shown. Both the DL-3DMM and the PCA-3DMM are able to model expressive faces but, nonetheless, our model has some advantages: 1) using the optimal λ value it introduces less noise in the resulting 3D model with respect to the PCA one; 2) if a smoother model is desired, the regularization value can be increased without sacrificing modeling ability. The PCA-3DMM, on the other hand, is not able to fit the expression properly in this case.

B. Cross-Dataset 3D Shape Reconstruction

We performed a cross-dataset fitting experiment using the FRGC dataset in addition to the BU-3DFE. The FRGC v2.0 includes 4,007 scans of 466 subjects acquired with frontal view from the shoulder level, with very small pose variations. About 60% of the faces have neutral expression, while the others show spontaneous expressions of disgust, happiness, sadness, and surprise. Scans are given as matrices of 3D points of size 480×640, with a binary mask indicating the valid points of the face. 2D RGB images of the face are also available and aligned with the matrix of 3D points. Ground-truth landmarks are not available in this dataset. To apply our alignment procedure, we first run the landmark detector in [48] to extract 68



(a) (b) (c)

Fig. 4. Reconstruction error on the BU-3DFE dataset: (a) 3D-2D fitting with front view; (b) 3D-2D fitting with side view; and (c) direct 3D-3D fitting. Each plot in the first row reports the errors for both DL- and PCA-based 3DMM as a function of the regularization parameter λ and for different number of components. The second row reports, for the best number of components, the effect of varying λ, while in the third row the effect of varying the number of components for the best λ value is shown. Standard deviation is also reported for each bar. (a) 3D-2D fitting, front view. (b) 3D-2D fitting, side view. (c) 3D-3D fitting.

points from the 2D images (the detection failed on just 6 images). Since 2D images and matrices of 3D points are aligned pixel-wise, the 2D landmarks position, plus 6 landmarks in the forehead of the face, can be transferred to 3D scans. Then, the alignment process described in Section III is applied. In order to have a meaningful alignment between the two datasets, the same partitioning described above has been applied to the BU-3DFE considering a subset of 68 out of the 83 landmarks available as ground truth and re-aligning the whole dataset.

In this experiment, the whole FRGC dataset was used to construct the average model $\hat{\mathbf{m}}$, the deformation components $\hat{\mathbf{d}}_i$, and the weights $\hat{\mathbf{w}}$, while all the models of the BU-3DFE have been used for test. The same experiment was performed considering the BU-3DFE as train and the FRGC for test. Reconstruction errors obtained for both DL- and PCA-based 3DMM shape fitting are reported in Fig. 7. It is possible to appreciate that when the FRGC is used for train, the reconstruction error is higher for both DL- and PCA-based 3DMM. A possible motivation for this is that, though the FRGC dataset contains about four times the number of identities of the BU-3DFE, it includes less intense expressions. Comparing the results of the DL- and PCA-based 3DMM, they are very close, even though DL obtains a slightly

Fig. 5. Example of the deformation obtained using single dictionary atoms. In column (a), the deformation heat-maps are reported; the models generated by applying different deformation magnitudes are shown in the other columns.

Fig. 6. 3DMM fitting examples with both DL- and PCA-based 3DMM for optimal or high regularization values. It is appreciable how our DL-3DMM both introduces less noise in the 3D models and retains its modeling ability even for high regularization values (face images from the CK+ dataset).

smaller error. On the other hand, when the BU-3DFE is used to learn the deformation components, the error decreases of about 2 mm. We explain this with the fact that adding more heterogeneous expression variations in the training permits the model to have a larger spectrum of deformations that ultimately result in more accurate reconstructions.

C. Facial Action Unit Detection and Emotion Recognition

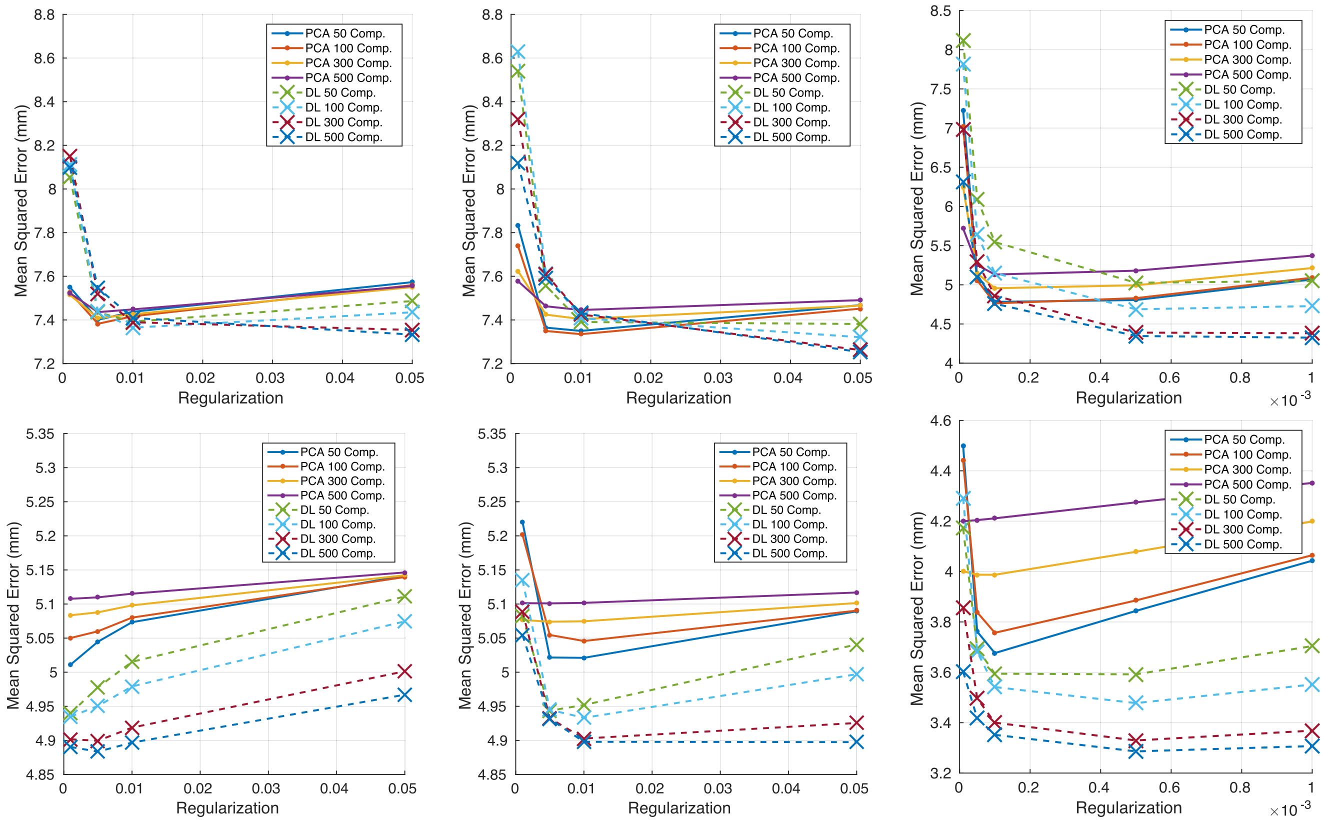
Using the pipeline of Section VI, in the following we report experimental results obtained by applying 3DMM to AU detection and emotion recognition.

State of the art methods for AU detection and emotion recognition [54], [60]–[62] have been evaluated and compared mainly on the *Extended Cohn-Kanade* (CK+) [57] and the *Facial Expression Recognition and Analysis* (FERA) [63] datasets. The CK+ dataset contains image sequences of posed and non-posed spontaneous expressions of 123 subjects (593 sequences in total). Each sequence has an average duration of about 20 frames, with the initial neutral expression varying up to a peak. The peak frame is AU-labeled, while an emotion label is associated to the entire sequence. The FERA dataset contains video sequences of 7 trained actors portraying 5 emotions. As in [54], [62], we used the training subset, which includes 87 videos ranging between 40 and 110 frames in length. Each frame is AU-labeled, while there is a single emotion label for the entire sequence. In both the datasets, the head pose is frontal in most of the sequences.

In the experiments, face images are described by LBP features [58], with a radius of 10px, following four different configurations:

1) *Dense grid*, DeGr: First, the face image is cropped. Then, eyes position is retrieved from landmark detection, and used to align the image to a common reference. In this phase, in-plane rotations are compensated. Finally, the image is resized to 200×200 pixels, and LBP descriptors are computed over 20×20 non overlapping patches;
2) *Landmarks*, LM: LBP descriptors are computed over patches centered in correspondence to 49 landmarks detected on the original image using the method in [48];
3) DL-(O) or PCA-(O): LBP descriptors are computed over patches localized by a subset of the vertices of the 3DMM, projected onto the original image;
4) DL-(F) or PCA-(F): LBP descriptors are computed over patches localized by a subset of the vertices of the 3DMM, projected onto the frontalized image.

The first two solutions do not use the 3DMM; the third and fourth, instead, perform local image description exploiting the localization provided by the 3DMM vertices. We experimentally found that a uniform subsampling of the vertices with step of 7 is the best balance between the face descriptor dimension and the patches overlap ratio. In fact, it is known that high dimensional



(a) (b) (c)

Fig. 7. Cross-dataset reconstruction errors obtained using FRGC for train and BU-3DFE for test (top) or vice versa (bottom). (a) 3D-2D fitting with frontal camera. (b) 3D-2D fitting with side camera. (c) Direct 3D-3D fitting. Each plot reports the errors for both DL- and PCA-based 3DMM as a function of the regularization parameter λ, and for different number of components. (a) 3D-2D fitting, front view. (b) 3D-2D fitting, side view. (c) 3D-3D fitting.

face descriptors and a large overlapping ratio between patches improve the effectiveness of the face description [64]. For each modality, we concatenate the LBP extracted from a face so as to form a unique descriptor, and reduce the descriptor dimensionality by applying PCA with a number of PCs that retain at least the 95% of variance.

AU detection: According to the experimental setup suggested in [57], [63], both for the CK+ and FERA datasets a *leave-one-subject-out* cross validation has been performed. For the CK+, only the neutral (first frame) and peak frames of each video sequence were used (the peak frame is the only one labeled). On the contrary, the FERA dataset comes with AUs labeled for each frame. However, not all the frames of a sequence have been used in the training phase since AUs are characterized mainly by an onset, a peak, and an offset phase. As suggested in [63], for each sequence, we consider the set of consecutive frames labeled with the peak label, and take its middle frame as corresponding to the peak phase.

Since the effect of each AU is limited to a portion of the face, accordingly to [63], AUs have been divided into *upper* and *lower* AUs corresponding to the upper half and lower half of the face, respectively. To train the SVMs, we used only the descriptors computed on points in the lower or upper part of the face, depending on which AU is considered. Each SVM is also trained independently, without accounting for the semantic relationships between different AUs (e.g., if the AU associated to the eyebrows raising is active, the AU associated to the eyebrows lowering cannot be active).

In Tables I and II, we report the AU detection results for the CK+ and FERA datasets, respectively. Detection performance is measured in terms of $F1$-score (i.e., the harmonic mean of *precision* and *recall*) and *Area Under the ROC Curve* (AUC). Three main facts emerge evidently: First, localizing the descriptors with either DL-3DMM or PCA-3DMM, rather than using the regular dense grid improves the results, since the alignment is more significant; Secondly, the greater number of points provided by the projected mesh allows the computation of more descriptors, which improves the performance; Lastly, the alignment and consistency of the image representation provided by our frontalization improves the discriminating power resulting in higher overall results. This behavior is more evident for the FERA dataset, which is more challenging than the CK+. Indeed, the continuous and spontaneous nature of the sequences included in the FERA dataset induces strong nuisances in the resulting feature descriptors. The alignment and consistency obtained with our representation, however, proved to be effective in reducing the complexity to be learned by the classifier, increasing the overall results on both CK+ and FERA.

For the comparison between using DL or PCA for 3DMM shape fitting, on the CK+ results are very close and this is in

TABLE I
AU DETECTION ON CK+

AU	$F1$-score						AUC					
	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
1	77.6	75.8	84.8	81.2	84.9	83.2	95.4	95.4	98.4	98.0	98.6	98.2
2	81.2	79.2	81.4	79.7	79.4	77.2	96.7	94.2	97.6	97.3	97.5	96.9
4	71.0	67.0	77.7	79.9	80.6	79.3	92.9	91.9	95.9	96.3	96.6	97.0
5	72.5	81.3	78.2	78.7	79.6	79.1	95.7	98.0	96.6	96.1	97.8	97.5
6	68.4	66.4	67.7	72.2	70.5	68.8	94.8	94.0	95.2	95.0	95.6	95.6
7	58.2	60.7	60.0	64.8	65.1	64.9	87.7	91.8	90.1	91.4	90.1	91.9
9	85.9	91.7	88.9	90.7	90.3	92.1	99.4	99.5	99.6	99.6	99.6	99.6
11	45.1	36.1	30.8	32.3	41.7	40.6	91.2	89.1	92.8	92.6	92.9	94.5
12	85.2	81.5	85.1	84.4	85.9	84.1	98.5	98.4	98.8	98.5	98.9	98.5
15	71.3	60.2	74.0	73.2	77.6	76.1	94.7	94.9	95.5	95.4	96.2	96.5
17	80.5	73.8	83.1	84.7	83.0	82.8	95.9	94.2	97.1	97.5	97.9	97.9
20	74.7	76.0	81.5	81.7	85.4	83.6	97.7	96.5	98.4	98.0	98.6	98.5
23	52.8	69.9	58.5	64.1	69.3	65.1	91.2	95.0	94.7	95.4	94.7	94.8
24	58.3	58.8	62.7	64.5	59.7	62.8	88.6	92.8	91.9	93.0	93.0	93.7
25	88.3	92.8	92.3	91.1	92.6	91.1	97.9	98.8	99.0	98.8	99.0	98.9
26	41.5	37.6	33.7	35.5	38.1	30.8	91.6	89.0	89.6	88.8	89.7	90.2
27	89.1	90.9	89.5	89.4	90.7	91.9	99.6	99.8	99.8	99.8	99.8	99.8
Avg.	75.3	75.1	78.2	78.9	**80.0**	78.8	95.3	95.5	96.7	96.7	97.0	**97.1**

Comparison of different feature extraction modalities. Results are reported in terms of $F1$-score and AUC. The average is weighted with respect to the number of positive instances, as indicated in [57].

TABLE II
AU DETECTION ON FERA

AU	$F1$-score						AUC					
	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
1	47.7	55.9	64.8	63.8	65.3	70.2	77.7	78.8	83.0	81.9	85.1	83.9
2	56.2	54.7	62.0	62.6	61.3	65.6	63.5	71.0	80.7	79.1	79.2	85.8
4	17.4	32.2	25.8	20.0	26.1	29.5	48.2	53.1	46.5	51.8	52.1	54.7
6	55.5	52.6	60.8	57.0	66.7	66.3	73.0	77.3	76.3	72.9	81.0	80.0
7	48.3	55.8	45.5	47.7	52.9	52.0	71.1	66.8	57.5	57.0	62.1	64.9
12	39.2	55.1	55.2	55.9	58.0	59.3	66.5	62.9	64.8	66.5	63.9	64.9
15	68.5	65.0	77.2	77.1	79.7	80.4	73.8	81.5	84.6	82.7	85.8	87.5
17	26.4	25.8	36.9	42.6	31.1	33.1	60.5	66.9	65.3	69.9	58.8	61.7
Avg.	44.9	49.6	53.5	53.4	55.1	**57.1**	66.8	69.8	69.8	70.2	71.0	**72.9**

Comparison of different feature extraction. Results are reported in terms of $F1$-score and AUC.

some way expected. In this dataset, for each sequence, we have that only the peak frame is AU labeled. Furthermore, the expressions shown are also rather exaggerated, as appreciable in the examples of Fig. 6. This makes the separation between the activation of different AUs somewhat easy and localizing the descriptors with sufficient precision becomes not crucial. This is proved by the fact that results on this dataset tend generally to saturate towards the maximum, with a rather small gap between baseline methods (DeGr and LM) and the 3DMM. The FERA dataset is instead much more challenging. The continuous and spontaneous nature of the sequences makes the gap between baseline methods (DeGr and LM) and the 3DMM increase significantly, supporting the usefulness of the latter. Finally, results show that DL performs better than PCA-3DMM on this dataset; this is mainly motivated by the fact that the face variations are more subtle and smooth and thus a better modeling improves the classification performance.

In Tables III and IV we provide a comparison with the state of the art in terms of average $F1$-score and AUC values. For the sake of completeness, results for the CjCRF method [61] on FERA are also reported, though they have been obtained by testing only on 260 frames out of the about 5000 total frames. Our method obtains comparable performance with respect to the state of the art on both datasets. Lower performance on the FERA dataset is likely due to the fact that our solution uses off-the-shelf descriptors and classifiers, and does not compensate

TABLE III
AU DETECTION ON CK+

Method	$F1$-score	AUC
IF [54]	76.6	91.3
Wang *et al.* [65]	**82.4**	96.7
CjCRF [61]	80.7	94.9
PCA-(F)	80.0	97.0
DL-(F)	78.8	**97.1**

Comparison with the state of the art. Results are reported in terms of $F1$-Score and AUC.

TABLE IV
AU DETECTION ON FERA

Method	$F1$-score	AUC
Wang *et al.* [65]	52.3	-
Data-Free [66]	52.6	-
IF [54]	59.0	**74.5**
DICA [62]	59.1	-
CjCRF [61]*	**59.6**	-
PCA-(F)	55.1	71.0
DL-(F)	57.1	72.9

Comparison with the state of the art. Results are reported in terms of $F1$-Score and AUC.

TABLE V
EMOTION RECOGNITION ON CK+

Emotion	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
Anger	97.6	99.0	98.6	98.8	98.9	99.4
Contempt	99.8	99.6	99.9	99.7	99.8	99.9
Disgust	99.2	97.3	97.3	93.9	99.6	99.7
Fear	99.9	99.9	99.9	99.9	99.9	99.9
Happiness	98.2	99.9	99.7	99.2	98.6	99.0
Sadness	98.8	98.9	99.2	99.0	98.8	98.8
Surprise	98.1	99.6	99.4	99.3	97.6	99.4
Avg.	98.8	99.1	99.2	98.6	99.1	**99.5**

Comparison of different feature extraction modalities. Results are reported in terms of AUC.

TABLE VI
EMOTION RECOGNITION ON FERA

Emotion	DeGr	LM	PCA-(O)	DL-(O)	PCA-(F)	DL-(F)
Anger	56.4	66.7	64.4	63.0	67.7	70.5
Fear	85.8	73.7	77.4	73.0	81.9	88.4
Joy	93.0	91.4	90.9	91.9	92.1	91.5
Relief	80.2	76.4	77.4	75.6	79.5	79.0
Sadness	81.1	78.0	81.0	80.7	86.2	81.5
Avg.	79.3	77.2	78.2	76.8	81.5	**82.2**

Comparison with the state of the art. Results are reported in terms of AUC.

TABLE VII
EMOTION RECOGNITION ON CK+

Emotion	IF [54]	PCA-(F)	DL-(F)
Anger	96.4	98.9	**99.4**
Contempt	96.9	99.8	**99.9**
Disgust	96.0	99.6	**99.7**
Fear	95.5	**99.9**	**99.9**
Happiness	98.9	98.6	**99.0**
Sadness	93.3	**98.8**	**98.8**
Surprise	97.6	97.6	**99.4**
Avg.	96.4	99.1	**99.5**

Comparison with the state of the art. Results are reported in terms of AUC.

TABLE VIII
EMOTION RECOGNITION ON FERA

Emotion	IF [54]	PCA-(F)	DL-(F)
Anger	**78.6**	67.7	70.5
Fear	85.5	81.9	**88.4**
Joy	**95.0**	92.1	91.5
Relief	**88.4**	79.5	79.0
Sadness	84.8	**86.2**	81.5
Avg.	**86.5**	81.5	82.2

Comparison with the state of the art. Results are reported in terms of AUC.

directly for the influence of the identity in the training as is explicitly done in [54], [62]. We believe that in this sense still there is enough room for improvements.

Emotion recognition: Data used for emotion recognition have some particular characteristics: as in the AU case, in the CK+ dataset each sequence has only two labels, one for the neutral and one for the peak frame; in the FERA dataset instead, each sequence is marked with a single label, representing the emotion of the entire sequence. For the CK+ dataset, emotion recognition is performed by considering the peak frames of each sequence in both the train and test sets; for FERA, we subsample each sequence and consider only 1 frame every 10.

In Tables V and VI, we report emotion recognition results obtained using the four feature extraction methods presented in Section VII-C. Consistent with the AU detection case, the results on CK+ are saturated with a small gap between the solutions that include the 3DMM and the others. However, the ones that exploit 3DMM and frontalization are the best performing. Results on FERA, instead, show that there is actually a tangible advantage in using the 3DMM for emotion recognition. From Table VI we can see that DL-(F) and PCA-(F) are, respectively, the best and the second best performing solutions, but DeGr performs better than DL-(O) and PCA-(O). This behavior can be explained considering that emotion recognition is based on the observation of the whole face appearance. In this case, localizing the descriptors precisely seems to become less important than having a consistent and pixel-wise aligned image representation.

In Tables VII and VIII, we report our results in terms of AUC in comparison with state of the art solutions, respectively, for the CK+ and FERA datasets. We observe that our solution outperforms the state of the art on the CK+ dataset, but scores lower performance than [54] on FERA. As for AU detection, this deficit of performance can be safely ascribed to the fact that differently from [54], we do not compensate the identity influence in the training.

VIII. DISCUSSION AND CONCLUSION

In this work, we proposed a dictionary learning based method for constructing a 3DMM, and we have shown its effectiveness on AU detection and facial emotion recognition. Compared to traditional methods for 3DMM construction based on PCA, our solution has the advantage of permitting more localized variations of the 3DMM that can better adapt to expressive faces. This capability to account for fine face deformations also depends on the inclusion in the training data of faces with large expression variability. This required us to develop a new method to establish a dense, point-to-point, correspondence between training faces. We also proposed an approach to effectively deforming the 3DMM, which includes pose estimation, regularized ridge-regression fitting, and frontalized image rendering. The comparative evaluation of the DL- with the PCA-based 3DMM shows a clear advantage of the DL based solution in terms of 3D reconstruction error. Also, we showed that the proposed framework opens the way to the application of 3DMM to facial expression analysis. In particular, we obtained effective results for AU detection and emotion recognition, even using off-the-shelf image face descriptors and machine learning methods.

A potential drawback of a 3DMM that includes expressive scans is the difficulty in discriminating between components modeling identity traits and components modeling facial movements. Further investigation would be useful to determine: 1) if more accurate vertex correspondences can be found by using different landmark detectors that induce more uniform partitioning of faces (which would also improve visual appearance of our models); 2) if an extended solution can be found that balances the tradeoff between the efficiency of fitting against greater precision; and 3) if deviations beyond shape can be accounted for in an extended 3DMM (for example by applying DL also to the texture component of faces).

References

[1] H. Li and G. Hua, "Hierarchical-PEP model for real-world face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 4055–4064.

[2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[3] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.

[4] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[6] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, Nov. 2013.

[7] S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh, "An approach for automatic pain detection through facial expression," in *Proc. Comput. Sci.*, vol. 84, 2015 pp. 99–106.

[8] D. J. Walger, T. P. Breckon, A. Gaszczak, and T. Popham, "A comparison of features for regression-based driver head pose estimation under varying illumination conditions," in *Proc. IEEE Int. Work. Comput. Intell. Multimedia Understand.*, Nov. 2014, pp. 1–5.

[9] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning personalized models for facial expression analysis and gesture recognition," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 775–788, Apr. 2016.

[10] T. Zhang *et al.*, "A deep neural network driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[11] E. Piątkowska and J. Martyna, "Spontaneous facial expression recognition: Automatic aggression detection," in *Proc. Int. Conf. Hybrid Artif. Intell. Syst.*, 2012, pp. 147–158.

[12] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2009, pp. 1–7.

[13] E. Pontikakis, C. Nass, J. N. Bailenson, L. Takayama, and M. E. Jabon, "Facial expression analysis for predicting unsafe driving behavior," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 84–95, Apr. 2011.

[14] A. Yuce, H. Gao, G. Cuendet, and J. P. Thiran, "Action units and their cross-correlations for prediction of cognitive load during driving," *IEEE Trans. Affective Comput.*, to be published.

[15] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniahand, and D. Dunaway, "A 3D morphable model learnt from 10,000 faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 5543–5552.

[16] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1701–1708.

[17] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 4838–4846.

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

[19] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model-based automatic 3D/4D facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1438–1450, Jul. 2016.

[20] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. ACM Conf. Comput. Graph. Interactive Techn.*, 1999, pp. 187–194.

[21] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[22] I. Masi, C. Ferrari, A. Del Bimbo, and G. Medioni, "Pose independent face recognition by localizing local binary patterns via deformation components," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 4477–4482.

[23] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3D face recognition with a morphable model," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recog.*, Sep. 2008, pp. 1–6.

[24] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Dictionary learning based 3D morphable model construction for face recognition with varying expression and pose," in *Proc. Int. Conf. 3D Vis.*, 2015, pp. 509–517.

[25] D. Shahlaei and V. Blanz, "Realistic inverse lighting from a single 2D image of a face, taken under unknown and complex lighting," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, May 2015, pp. 1–8.

[26] L. Zhang *et al.*, "Image-driven re-targeting and relighting of facial expressions," in *Proc. Comput. Graph. Int.*, 2005, pp. 11–18.

[27] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 641–650, 2003.

[28] F. C. Staal *et al.*, "Describing Crouzon and Pfeiffer syndrome based on principal component analysis," *J. Cranio-Maxillofacial Surgery*, vol. 43, no. 4, pp. 528–536, 2015.

[29] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li, "Discriminative 3D morphable model fitting," in *Proc. IEEE 11th Int. Conf. Autom. Face Gesture Recog.*, May 2015, vol. 1, pp. 1–8.

[30] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveillance*, Sep. 2009, pp. 296–301.

[31] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[32] A. Patel and W. A. P. Smith, "3D morphable face models revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1327–1334.

[33] D. Cosker, E. Krumhuber, and A. Hilton, "A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2296–2303.

[34] A. Brunton, T. Bolkart, and S. Wuhrer, "Multilinear wavelets: A statistical shape space for human faces," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 297–312.

[35] M. Lüthi, C. Jud, T. Gerig, and T. Vetter, "Gaussian process morphable models," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1603.07254

[36] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, pp. 986–993.

[37] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3539–3545.

[38] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 146–155.

[39] M. Grupp, P. Kopp, P. Huber, and M. Rätsch, "A 3D face modelling approach for pose-invariant face recognition in a human-robot environment," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1606.00474

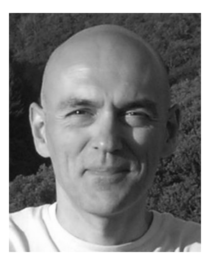
[40] G. Hu *et al.*, "Face recognition using a unified 3D morphable model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 73–89.
[41] S. Ramanathan, A. Kassim, Y. V. Venkatesh, and W. S. Wah, "Human facial expression recognition using a 3D morphable model," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 661–664.
[42] H. Ujir and M. Spann, *Facial Expression Recognition Using FAPs-Based 3DMM*. Dordrecht, The Netherlands: Springer, 2013, pp. 33–47.
[43] D. Cosker, E. Krumhuber, and A. Hilton, "Perceived emotionality of linear and non-linear AUs synthesised using a 3D dynamic morphable facial model," in *Proc. Facial Anal. Animation*, 2015, pp. 7:1–7:1.
[44] D. Cosker, E. Krumhuber, and A. Hilton, "Perception of linear and nonlinear motion properties using a FACS validated 3D facial model," in *Proc. 7th Symp. ACM Appl. Perception Graphics Vis.*, 2010, pp. 101–108.
[45] P. Huber, P. Kopp, M. Rätsch, W. J. Christmas, and J. Kittler, "3D face tracking and texture fusion in the wild," in *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1605.06764
[46] R. Kimmel and J. A. Sethian, "Computing geodesic paths on manifolds," *Proc. Nat. Academy Sci.*, vol. 95, no. 15, pp. 8431–8435, 1998.
[47] X. Lu and A. K. Jain, "Deformation modeling for robust 3D face matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1346–1357, Aug. 2008.
[48] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1867–1874.
[49] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 532–539.
[50] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris, "3D facial landmark detection under large yaw and expression variations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1552–1564, Jul. 2013.
[51] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, Apr. 2006, pp. 211–216.
[52] L. G. Farkas, *Anthropometry of the Head and Face*. New York, NY, USA: Raven Press, 1994.
[53] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Mach. Learning*, 2009, pp. 689–696.
[54] F. De la Torre *et al.*, "Intraface," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, May 2015, vol. 1, pp. 1–8.
[55] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Effective 3D based frontalization for unconstrained face recognition," in *Proc. 23rd Int. Conf. Pattern Recog.*, 2016, pp. 1047–1052.
[56] C.-H. Hjortsjö, *Man's Face and Mimic Language*. Lund, Sweden: Studen Litteratur, 1969.
[57] P. Lucey *et al.*, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2010, pp. 94–101.
[58] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recog.*, vol. 29, no. 1, pp. 51–59, 1996.
[59] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Workshop Face Recog. Grand Challenge Exp.*, Jun. 2005, pp. 947–954.
[60] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3515–3522.
[61] Y. Wu and Q. Ji, "Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 3400–3408.
[62] C. Georgakis, Y. Panagakis, and M. Pantic, "Discriminant incoherent component analysis," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2021–2034, May 2016.
[63] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 42, no. 4, pp. 966–979, Aug. 2012.
[64] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3025–3032.
[65] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3304–3311.
[66] Y. Li, J. Chen, Y. Zhao, and Q. Ji, "Data-free prior model for facial action unit recognition," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 127–141, Apr./Jun. 2013.

Claudio Ferrari received the M.S. degree (*cum laude*) in computer engineering from the University of Florence, Florence, Italy, in 2014, and is currently working toward the Ph.D. degree at the University of Florence under the supervision of Prof. A. Del Bimbo.

He is currently a Research Fellow at the Media Integration and Communication Center, University of Florence. His research interests include pattern recognition and computer vision mostly focused on the 2D/3D unconstrained face recognition problem.

Giuseppe Lisanti received the Ph.D. degree in computer engineering from the University of Florence, Florence, Italy, in 2012.

He is a Postdoctoral Researcher with the Media Integration and Communication Center, University of Florence. His main research interests include computer vision and pattern recognition, specifically for person detection and tracking, person reidentification, 2D, and 3D face recognition.



Stefano Berretti (M'07–SM'16) is currently an Associate Professor with the University of Florence, Florence, Italy. His research interests focus on image content modeling, indexing and retrieval, 3D computer vision for face based biometrics, human emotion and behavior understanding.

Prof. Berretti is the Information Director of the *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Alberto Del Bimbo is a Full Professor of Computer Engineering and the Director of the Media Integration and Communication Center, University of Florence, Florence, Italy. His scientific interests include multimedia information retrieval, pattern recognition, image and video analysis and natural human–computer interaction.

Prof. Del Bimbo is IAPR Fellow, an Associate Editor of several leading journal in the area of pattern recognition and multimedia, and the Editor-in-Chief of the *ACM Transactions on Multimedia Computing, Communications, and Applications*. He was also the recipient of the prestigious SIGMM 2016 Award for Outstanding Technical Contributions to Multimedia Computing, Communications and Applications.