3D Face Reconstruction from RGB-D Data by Morphable Model to Point Cloud Dense Fitting

Claudio Ferrari, Stefano Berretti, Pietro Pala and Alberto Del Bimbo
Media Integration and Communication Center, University of Florence, Florence, Italy

Keywords: 3DMM Construction, 3DMM Fitting, 3D Face Analysis.

Abstract: 3D cameras for face capturing are quite common today thanks to their ease of use and affordable cost. The depth information they provide is mainly used to enhance face pose estimation and tracking, and face-background segmentation, while applications that require finer face details are usually not possible due to the low-resolution data acquired by such devices. In this paper, we propose a framework that allows us to derive high-quality 3D models of the face starting from corresponding low-resolution depth sequences acquired with a depth camera. To this end, we start by defining a solution that exploits temporal redundancy in a short-sequence of adjacent depth frames to remove most of the acquisition noise and produce an aggregated point cloud output with intermediate level details. Then, using a 3DMM specifically designed to support local and expression-related deformations of the face, we propose a two-steps 3DMM fitting solution: initially the model is deformed under the effect of landmarks correspondences; subsequently, it is iteratively refined using points closeness updating guided by a mean-square optimization. Preliminary results show that the proposed solution is able to derive 3D models of the face with high visual quality; quantitative results also evidence the superiority of our approach with respect to methods that use one step fitting based on landmarks.

1 INTRODUCTION

Recent advances in 3D scanning technologies make it possible to acquire registered RGB and depth frames at affordable cost. The availability of depth and RGB data greatly simplifies the design of video analysis modules for object detection and segmentation. In fact, discontinuities in the depth domain highlight the presence of object boundaries that can be difficult to detect in the RGB domain especially in the case of background clutter and/or uneven illuminating conditions. However, a common trait of these low-cost RGB-D cameras—including the Microsoft Kinect, the Asus Xtion and the Intel RealSense depth cameras—is that depth data of individual frames are badly affected by noise that prevents the accurate reconstruction of the 3D geometry of the observed scene. This is particularly true if the task of reconstructing the 3D geometry of non-planar object surfaces is targeted, such as the case of reconstructing the 3D geometry of faces. In the last few years, several 3D face reconstruction approaches have been proposed, also in truly uncooperative, *in the wild*, conditions (Booth et al., 2018). However, a common trait of these approaches is that reconstruction of the 3D face model from data observed in a generic image or video frame is finalized to reproduce realistic renderings of the observed face in a different pose (*e.g.*, frontal pose) for the purpose of boosting the accuracy of person or facial expression recognition. In these solutions, smoothness of the reconstructed 3D face model is privileged with respect to fitting to the actual 3D geometry of the face. Indeed, this yields pleasant and realistic face renderings, but may prove inadequate to provide a precise reconstruction of the 3D geometry of the face and its deformations in the presence of voluntary and involuntary expressions.

Motivated by these premises, in this paper we propose a novel face modeling approach that starting from an RGB-D low-resolution sequence of the face is capable of reconstructing an accurate 3D face model over time also in the presence of facial expressions and generic facial deformations. This opens the way to application scenarios that aim to analyse the face deformations at a fine scale so as to understand the 3D local face variations with respect to a neutral model. For example, this can have applications in face rehabilitation, where a subject at home (*e.g.*, a patient recovering after a stroke or a face surgery) can be instructed to perform some facial deformations in front

Ferrari, C., Berretti, S., Pala, P. and Bimbo, A.
3D Face Reconstruction from RGB-D Data by Morphable Model to Point Cloud Dense Fitting.
DOI: 10.5220/0007521007280735
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 728-735
ISBN: 978-989-758-351-3

of a PC. In such scenario, 3D high-resolution scanners cannot be employed due to their size and cost, and RGB data alone cannot provide the required level of accuracy. In our proposed solution, we start by the idea of using a 3DMM that also includes modes of deformation associated to facial expression variations and is specifically designed to account for local changes of the face. This is possible thanks to a dictionary learning framework that gives the atoms of the dictionary the capability of producing local deformations of the average model of the face. The model is then fit to a point cloud as captured by a low-resolution scanner (*e.g.*, Kinect) through a coarse-to-fine solution: the initial fitting is driven by the correspondence of a set of landmarks; the coarsely deformed model is then refined by an iterative closest points reassignment that minimizes the mean square error between corresponding points in an ICP like manner. Preliminary results have been obtained by measuring the reconstruction error between the 3D models derived by fitting the 3DMM to the target scans of two large face datasets and the target scans themselves used as ground truth. Results are univocal in indicating superior accuracy for the proposed solution with respect to using a one step 3DMM fitting based on landmarks.

The rest of the paper is organized as follows: In Section 2, we summarize the previous works on 3DMM construction and fitting that are more close to our proposal; In Section 3, we first report on the proposed method for 3DMM construction; Then, the way the 3DMM is fit to depth scans is detailed; The denoising operation applied to low-resolution sequences of face scans before 3DMM fitting is introduced in Section 4; Experimental results are reported in Section 5; Section 6 concludes the paper by reporting discussion and directions for future work.

2 RELATED WORK

In general, methods capable of reconstructing a 3D model of the face from low-resolution depth data can be categorized as either *driven by the data* or based on *fitting a 3D (morphable) model*.

Methods in the first category build a 3D face model by integrating tracked live depth images into a common final 3D model. For example, a method to produce laser scan quality 3D face models from a freely moving user with a low-cost, low-resolution depth camera was proposed in (Hernandez et al., 2012). The model is initialized with the first depth image, and then each subsequent cloud of 3D points is registered to the reference using a GPU implementation of the ICP algorithm. This registration is robust in that it rejects poor alignment due to facial expressions, occlusions, or a poor estimation of the transformation. Temporal and spatial smoothing of the successively incremented model is performed thanks to the introduction of the *Bump Images* framework to parameterize the 3D facial surface in cylindrical coordinates. One evident limitation of this approach is that it is not capable of reconstructing expressive models of the face. In (Newcombe et al., 2011), the Kinect Fusion approach was proposed to fuse all of the depth data streamed from a Kinect sensor into a single global implicit surface model of the observed scene in real-time. To this end, the current pose of the sensor is obtained by tracking the live depth frame relative to the global model using a coarse-to-fine ICP algorithm, which uses all of the observed depth data available. The Kinect Fusion method was developed for a fixed scene; in (Izadi et al., 2011) the method was extended by considering dynamic actions of the foreground. Though the Kinect Fusion approach is general, its application to 3D face reconstruction results into models that reduce the noise with respect to individual frames, but still show a quite visible gap with respect to high-quality scans. In (Anasosalu et al., 2013), a method is presented for producing an accurate and compact 3D face model in real-time using an RGB-D sensor like the Kinect camera. To this end, after initialization, Bump Images are updated in real time by using every RGB-D frame with respect to the current viewing direction and head pose; these latter are estimated using a frame-to-global-model registration strategy. Though this method takes a live sequence of RGB-D images streamed from a fixed consumer RGB-D sensor with unknown head pose, it is assumed the relative movement of the head between two successive frames to be small, and that the facial expression does not change during reconstruction. The work in (Zollhöfer et al., 2014) presents a combined hardware/software solution for marker-less reconstruction of non-rigidly deforming objects with arbitrary shape. First, a smooth template model of the subject as he/she moves rigidly is scanned. This geometric surface prior is used to avoid strong scene assumptions, such as a kinematic human skeleton or a parametric shape model. Next, a GPU pipeline performs non-rigid registration of live RGB-D data to the smooth template using an extended non-linear as-rigid-as-possible framework. High-frequency details are fused onto the final mesh using a linear deformation model. However, this solution relies on a specific stereo matching algorithm to estimate real-time RGB-D data. Other solutions were specifically designed for faces with no expressions in constrained

scenarios (Berretti et al., 2014; Bondi et al., 2016).

Methods in the second category, *i.e.*, methods that use a 3DMM to reconstruct a face model from depth data, are few. In fact, in the most practiced solutions a 3DMM is fit to a 2D RGB target image (Blanz and Vetter, 2003; Ferrari et al., 2017), with applications that span from face rendering and relighting, to face and facial expression recognition. In this work, instead, we are interested to the particular case where the target is a 3D low-resolution face scan as can be acquired by a Kinect-like camera. The method described in (Zollhöfer et al., 2011) employed a 3DMM that is fit to the depth images obtained from an RGB-D camera. The template mesh and the incoming frame are aligned using features detected in the RGB image as a coarse alignment step. The template is then aligned non-rigidly to the incoming frame, and the 3DMM is fit to the template. Unfortunately, this approach produces results that are biased towards the template. The work of (Kazemi et al., 2014) contributes a real time method for recovering facial shape and expression from a single depth image. The output is the result of minimizing the error in reconstructing the depth image, achieved by applying a set of identity and expression blend shapes to the model. A discriminatively trained prediction pipeline is used that employs random forests to generate an initial dense, but noisy correspondence field. Then, a fast ICP-like approximation is exploited to update these correspondences, allowing a quick and robust initial fit of the model. The model parameters are then fine tuned to minimize the true reconstruction error using a stochastic optimization technique.

However, none of these solutions can reconstruct fine details of expressive faces using a 3DMM.

3 3D MORPHABLE FACE MODEL

The work in (Blanz and Vetter, 1999) 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. The 3DMM was further refined into the Basel Face Model in (Paysan et al., 2009) with several other subsequent evolutions (Patel and Smith, 2009; Booth et al., 2016). Expressive scans were not part of the training in all the solutions above. Indeed, two aspects have a major relevance in characterizing the different methods for 3DMM construction: (1) the human face variability captured by the model, which directly depends on the number and heterogeneity of training scans; (2) the capability of the model to account for facial expressions; also this feature directly derives from the presence of expressive scans in the training. One of the few 3DMM in the literature that exposes both these features is the Dictionary Learning based 3DMM (DL-3DMM) proposed in (Ferrari et al., 2017). Since we mainly develop on this model, below we first describe the peculiar features that make the DL-3DMM suitable for our purposes, then we focus on the proposed fitting procedure.

DL-3DMM Construction. The first problem to be solved in the construction of a 3DMM is the selection of an appropriate set of training data. This should include sufficient variability in terms of ethnicity, gender, age of the subjects so as to include a large variance in the data. Apart for this, the most difficult aspect in preparing the training data is the need to provide dense, *i.e.*, vertex-by-vertex, alignment between the 3D scans. Differently from works in the literature that either use optical-flow (Blanz and Vetter, 1999) or the non-rigid ICP algorithm (Paysan et al., 2009), the dense alignment of the training data for the DL-3DMM was obtained with a solution based on face landmarks detection. These landmarks are then used for partitioning the face into a set of non-overlapping regions, each one identifying the same part of the face across all the scans. Re-sampling the internal of the region based on its contour, a dense correspondence is derived region-by-region and so for all the face. Such method showed to be robust also to large expression variations as those occurring in the Binghamton University 3D facial Expression (BU-3DFE) database (Yin et al., 2006). This dataset was used in the construction of the DL-3DMM.

Once a dense correspondence is established across the training data, these are used to estimate a set of M deformation components $\mathbf{C}_i$, usually derived by PCA, that will be linearly combined to generate novel shapes $\mathbf{S}$ starting from an average model $\mathbf{m}$:

$$\mathbf{S} = \mathbf{m} + \sum_{i=1}^{|M|} \mathbf{C}_i \alpha_i \, . \tag{1}$$

In the DL-3DMM, a dictionary of deformation components is learned by exploiting the *Online Dictionary Learning for Sparse Coding* technique (Mairal et al., 2009). Learning is performed in an unsupervised way, without exploiting any knowledge about the data (*e.g.*, identity or expression labels). The average model is deformed using the dictionary atoms $\mathbf{D}_i$ in place of $\mathbf{C}_i$ in Eq. (1). More details on the dictionary learning procedure can be found in (Ferrari et al., 2017). The average model $\mathbf{m}$, the dictionary $\mathbf{D}$ and $\mathbf{w}$, constitute the DL-3DMM.

3DMM Fitting. The 3DMM was originally designed with the goal of reconstructing the 3D shape of a face from single images (Blanz and Vetter, 2003); the large number of different techniques to fit the 3DMM to a face image developed later on can be divided in two main categories: *analysis-by-synthesis*, and *geometric based*. Methods in the former category perform a complex iterative procedure aimed at generating a synthetic image as similar as possible to the input one, optimizing with respect to the 3DMM (shape and texture) and rendering (*e.g.*, camera or illumination) parameters. Despite their complexity, the resulting reconstructions are rather accurate. Nonetheless, given a textured rendering, it is hard to discern if the retrieved shape resembles the real geometry of the face; this because the same rendering might be the result of different combinations of the many involved parameters. Alternatively, methods in the geometric-based category try to deform the 3DMM so as to match some geometrical features detected on the image, like facial landmarks or edges (Bas et al., 2016). These approaches exploit the fact that human faces are composed of muscles—hence facial movements involve an extended surface rather than a single point—that are constrained to fixed anthropometric proportions and limited variability. Thus, when trying to deform the 3DMM to fit a set of sparse landmarks, the surrounding surfaces will smoothly follow the deformation in a statistically plausible way. This motivates the coarse reconstruction of the shape of the whole face based only on few control points. Obviously, the resulting reconstruction will be a coarse, but smooth approximation of the real surface.

The proposed method attempts to fill this gap; it builds upon the geometric approaches and extends the fitting based on facial landmarks to a whole point cloud. This extension implies a 3D scan corresponding to the face image to be available, which changes the context and the objective. In this configuration, the goal becomes deforming a generic face shape to match a target one, both represented as point clouds; indeed, the problem can be seen as the non-rigid registration of point clouds, which is a well-known problem in computer vision for which many solutions have been proposed throughout the years (Chui and Rangarajan, 2000; Amberg et al., 2007; Myronenko and Song, 2010). All the approaches addressing the problem, however, are intended to work with generic point clouds representing arbitrary objects, while in this case the problem is bounded to human faces. The main difficulty is that faces are highly deformable objects, which often makes such approaches fail in matching the two shapes. On the opposite, we can exploit this prior knowledge to leverage a statistical tool such as the 3DMM to bound the deformation.

The proposed approach, first performs a similarity transformation to map the target shape into the average model space, accounting for 3D rotation, translation and scale (*SimilarityTransform* in Algorithm 1). This is achieved by means of a set of 49 landmarks $\mathbf{L}_t \in \mathbb{R}^{49\times 3}$, which are detected on the face image and back-projected to the mesh. Depending on the data, the association method to be applied might change.

To initialize the approach and account for large shape differences that might impair the subsequent steps, we apply the DL-3DMM fitting using the landmarks similarly to (Ferrari et al., 2015). The average model $\mathbf{m} \in \mathbb{R}^{p\times 3}$ is deformed on the target shape $\widetilde{\mathbf{t}} \in \mathbb{R}^{k\times 3}$ minimizing the Euclidean distance of the landmarks, whose indices on $\mathbf{m}$ are indicated as $\mathbf{I}_l \in \mathbb{N}^{49}$ (*LandmarkFitting* in Algorithm 1). Differently from (Ferrari et al., 2015), the fitting is performed directly in the 3D space and projection on the image plane is avoided. The deformation coefficients α are retrieved using the dictionary atoms $\mathbf{d}_i$ as:

$$\min_{\alpha} \left\| \mathbf{L}_t - \mathbf{m}(\mathbf{I}_l) - \sum_{i=1}^{|\mathbf{D}|} \mathbf{d}_i(\mathbf{I}_l)\alpha_i \right\|_2^2 + \lambda \left\| \alpha \circ \hat{\mathbf{w}}^{-1} \right\|_2 . \quad (2)$$

In the equation, $\mathbf{d}_i(\mathbf{I}_l)$ indicates that, for each dictionary atom, only the elements associated to the vertices corresponding to the landmarks are involved in the minimization. The solution is found in closed form and the average model $\mathbf{m}$ is deformed using Eq. (1) to obtain an initial estimate $\hat{\mathbf{m}}$.

Then, we perform a rigid ICP registration between $\hat{\mathbf{m}}$ and $\widetilde{\mathbf{t}}$ to refine the alignment, and compute the per-vertex distance between the two meshes. We subsequently associate each vertex of $\hat{\mathbf{m}}$ to its nearest neighbor in $\widetilde{\mathbf{t}}$, obtaining a re-parametrization (*VertexAssociation* in Algorithm 1) of p indices of $\widetilde{\mathbf{t}}$. Note that $k \neq p$ in general, thus a vertex of $\hat{\mathbf{m}}$ might be associated with multiple vertices of $\widetilde{\mathbf{t}}$; even if $p = k$, this can still happen because of points that share the same nearest neighbor. Once the association is done, the DL-3DMM is fit minimizing the Euclidean distance between each pair of associated points, using a least squares solution. The fitting method reported in (Ferrari et al., 2015) uses a regularized formulation (Eq. (2)), which is necessary to avoid uncontrolled deformations. In our case, we use all the vertices to fit the target shape; thus the usefulness of the regularization becomes marginal. The minimization of Eq. (2) becomes:

$$\min_{\alpha} \left\| \widetilde{\mathbf{t}} - \hat{\mathbf{m}} - \sum_{i=1}^{|\mathbf{D}|} \mathbf{d}_i\alpha_i \right\|_2^2 . \quad (3)$$

The procedure is repeated until the error between subsequent iterations is lower than a threshold τ or the

Algorithm 1: Point Cloud Fitting (PCF).

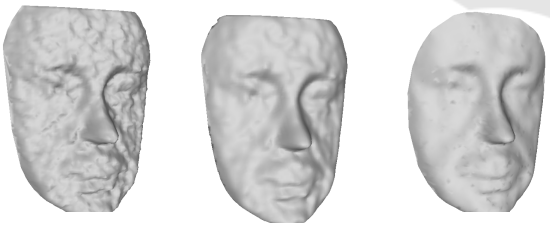
Input: Average Model $\mathbf{m}$, Dictionary $\mathbf{D}$, Weights $\mathbf{w}$, Target Shape $\mathbf{t}$, Landmarks $\mathbf{L}_t, \mathbf{m}(\mathbf{I}_l)$, Error Threshold τ, Iterations Limit $MaxIter$
Output: Deformed Model $\hat{\mathbf{m}}$

$\widetilde{\mathbf{t}}$ = SimilarityTransform($\mathbf{L}_t, \mathbf{t}, \mathbf{m}(\mathbf{I}_l)$);
$\hat{\mathbf{m}}$ = LandmarkFitting($\mathbf{L}_t, \mathbf{m}(\mathbf{I}_l), \mathbf{D}, \mathbf{w}$);
$i = 0$;
while $i < MaxIter \parallel err > \tau$ **do**
  ICP($\widetilde{\mathbf{t}}$,$\hat{\mathbf{m}}$);
  $\widetilde{\mathbf{t}}$ = VertexAssociation($\widetilde{\mathbf{t}}$,$\hat{\mathbf{m}}$);
  $\hat{\mathbf{m}}$ = ShapeFitting($\widetilde{\mathbf{t}}, \hat{\mathbf{m}}, \mathbf{D}, \mathbf{w}$);
  err = ComputeEuclideanError($\widetilde{\mathbf{t}}$,$\hat{\mathbf{m}}$);
  $i = i + 1$

maximum number of iterations if reached. Algorithm 1 reports the pseudo-code of the proposed Point Cloud Fitting procedure (PCF).

4 DEPTH DATA DENOISING

Low cost RGB-D scanners, such as the Kinect, can acquire multimodal video streams consisting of registered RGB and depth data at approximately 30fps. However, since the depth data are badly affected by noise, a pre-processing is needed for noise reduction and reconstruction of a regularized surface that approximates the 3D geometry of the observed object (*i.e.*, a face in our domain of interest). Depth data are



(a) (b) (c)

Figure 1: Effect of local regression filtering for depth noise removal: (a) raw depth scan affected by noise as provided by the Kinect; (b) depth data after smoothing through a Laplacian operator; (c) depth scan after local regression filtering (points closer than 9cm to the nose tip are retained).

regarded as z values of a function on the (x, y) plane, perpendicular to the line of sight of the scanner. Then, regularization is performed through the local regression scheme described in (Cleveland, 1979): estimation of the true depth value at point $\mathbf{p}_i = (x_i, y_i, z_i)$ is accomplished by fitting a low-dimensional polynomial to the nearest neighbors $\mathcal{N}(\mathbf{p}_i)$ of $\mathbf{p}_i$. Operatively, the cardinality of $\mathcal{N}(\mathbf{p}_i)$ is controlled through a *smoothing parameter* $\gamma \in (0, 1)$. Large values of γ produce smooth regression functions that wiggle the least in response to fluctuations in the data. The smaller γ is, the closer the regression function will conform to the data, thus yielding poor robustness to noise. The noise free estimate of point $\mathbf{p}_i$ is obtained by computing a least squares bilinear fit on pairs $(x_j, y_j) \mapsto (x_j, y_j, z_j), \forall (x_j, y_j, z_j) \in \mathcal{N}(\mathbf{p}_i)$. Figure 1 shows the effect of the application of the local regression module for noise reduction.

5 EXPERIMENTAL RESULTS

In the following, we report on the evaluation of the proposed approach for accurate 3D face reconstruction using the DL-3DMM and the PCF procedure. The experiments are conceived to demonstrate that multimodal RGB and depth data, even if affected by noise and provided by low resolution scanners, can be processed through the PCF procedure to reconstruct the 3D face shape. Furthermore, fitting the morphable model through PCF yields more accurate 3D reconstruction compared to fitting the model to the landmarks.

In order to quantitatively evaluate the reconstruction accuracy of the proposed approach, we identified two publicly available datasets of 3D facial scans, namely, the Binghamton University 3D Facial Expression database (BU-3DFE) (Yin et al., 2006) and the Face Recognition Grand Challenge database (FRGC) (Phillips et al., 2005).

The BU-3DFE dataset has been largely employed for 3D expression/face recognition; it 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 (2500 scans in total). The subjects are distributed across different ethnic groups or racial ancestries. This dataset has been used to train the DL-3DMM and a fully registered version of 1779 out of the 2500 scans is available.

The FRGC dataset is composed of 466 individuals, for a total of 4007 scans collected in two separate sessions. Approximately, the 60% of such are in neutral expression, while the others show spontaneous expressions. For the experiments, we used the "fall2003" session, comprising 1729 scans. In the following, we first present and discuss results on BU-3DFE and FRGC, then for some Kinect scans. For all the reported experiments, the regularization term λ of Eq. 2 has been fixed to 0.01; the error threshold τ and the maximum number of iterations in Algorithm 1

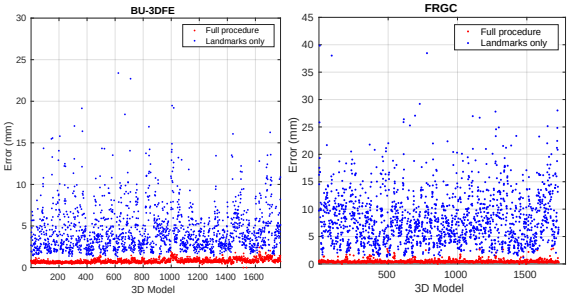


Figure 2: Comparison between landmark fitting (FL) with blue color, and full point cloud fitting (FL+PCF) with red color. BU-3DFE (left) and FRGC (right).

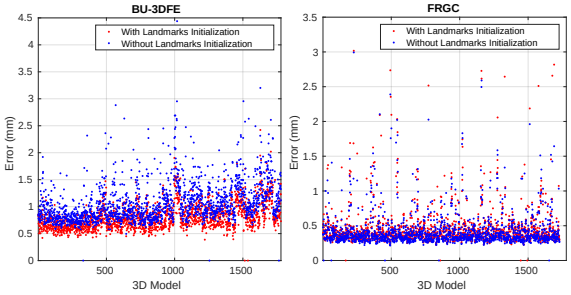


Figure 3: Comparison either using the landmark initialization or not. BU-3DFE (left) and FRGC (right).

have been fixed to 0.001 and 50 respectively.

Reconstruction Accuracy by Nearest Neighbor Association. In this experiment, we evaluate the accuracy of the vertices association by computing, for each vertex of the deformable model, its distance to the closest vertex of the 3D point cloud. It should be noticed that such nearest neighbor point association might not fully reflect the correctness of the reconstruction. Consider the case in which the 3DMM fails to fit an open mouth; the distance between the two models should be ideally large, while it is likely that the error resulting from a nearest neighbor search will be small (or at least, smaller). To highlight the potential of the proposed solution, values of the mean accuracy are computed with reference to two distinct cases: *i*) DL-3DMM fitting based on facial landmarks (FL), and *ii*) DL-3DMM fitting based on facial landmarks and PCF (FL+PCF). The mean error and standard deviation across all the models for the BU-3DFE and FRGC datasets are reported in Table 1.

Table 1: Reconstruction error (in mm) for the landmark fitting (FL) and full point cloud fitting (FL+PCF).

Dataset	FL	FL+PCF
BU-3DFE	4.507 ± 2.809	0.802 ± 0.235
FRGC	8.272 ± 4.939	0.455 ± 0.282

In Fig. 2 accuracy values are reported at a higher level of detail, showing for each model of the dataset (both BU-3DFE and FRGC are considered) the error obtained by using FL (in blue color), and the error obtained by using FL+PCF (in red color). Results demonstrate a noticeable improvement is obtained with the proposed procedure. To provide a qualitative yet representative description of the accuracy of 3D face reconstruction, Fig. 4 reports some heatmaps obtained by encoding with a chromatic value the error associated with each vertex of the reconstructed model; it can be appreciated how the real surface is accurately reconstructed and fine details of regions where landmarks are missing are accurately replicated, while maintaining a smoother surface. The higher accuracy of the proposed solution is demonstrated by the presence of large regions with blue/cyan colors (low error values) for models reconstructed using FL+PCF compared to regions with red/yellow colors (high error values) for models reconstructed using FL.

Landmarks Initialization. Detecting facial landmarks on face images (or 3D data) is itself a challenging task, which is not exempt from failures. In order to assess the robustness of our approach to the initialization procedure, we compared the final accuracy in case the method is whether initialized or not. From Fig. 3, we can observe that the approach is rather robust and still can accurately reconstruct the shape without landmarks initialization. For some models, the error increases significantly; this is the case of models with large topological differences, *e.g.*, open mouth, which are more difficult to handle if a semantic association is not available. Note that this happens mostly in the BU-3DFE, while for the FRGC dataset, such behavior happens in both the cases. This is motivated by the fact that, as above mentioned, the landmark detection can fail; if this happens, the initialized shape $\hat{\mathbf{m}}$ might be farther from the ground truth with respect to the average model $\mathbf{m}$.

Kinect Data Reconstruction. Our approach has been experimented also on Kinect data, processed as expounded in Sect. 4. Because of the lack of datasets containing RGB-depth pairs, we collected a few sequences where to qualitatively test our method on, while quantitative and statistically meaningful results were presented in the previous sections. The average error obtained in such sequences is $35,89$ and $1,70$ mm, respectively, for the FL and FL+PCF, demonstrating the effectiveness of the approach even for low-resolution data. From Fig. 4, we can appreci-

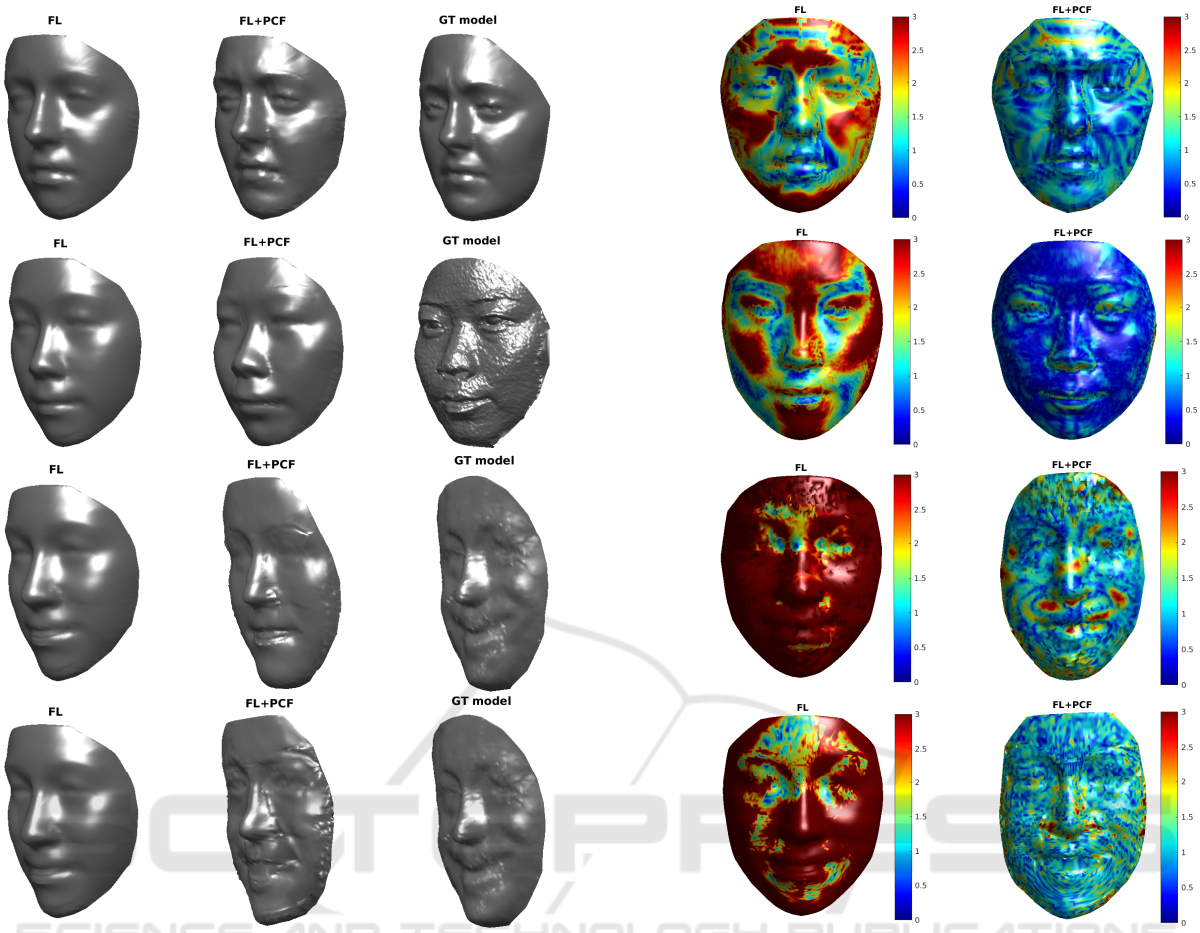


Figure 4: Qualitative comparison between DL-3DMM fitting based on facial landmarks (FL) and DL-3DMM fitting based on facial landmarks and PCF (FL+PCF). Three leftmost columns: reconstructions and ground truth (GT) models; two rightmost columns: error heat-maps. From top to bottom, BU-3DFE, FRGC, Kinect models (last two rows).

ate that the sole landmarks were not sufficient to reproduce the real shape, *e.g.*, the nose, but it could only coarsely capture the expression. A critical aspect of dealing with low-resolution data is the nearest neighbor association; each collected Kinect scan $\mathbf{k}$ has about 3K vertices, while the 3DMM has 6704, almost the double. If the association is performed with respect to the average model $\mathbf{m}$, then each vertex of $\mathbf{k}$ will be associated, on average, to 2 vertices of $\mathbf{m}$. The other way around, if we perform the association with respect to $\mathbf{k}$, not all the vertices of $\mathbf{m}$ will have a mated point. In the former case, the resulting reconstruction will be more accurate, but some noise due to the redundancy in the points association will be introduced (Fig. 4, bottom row); in the latter, the reconstruction would be smoother, but less accurate (Fig. 4, third row). Moreover, since we do not have control on the whole point cloud, the fitting needs to be performed using the regularization term, *i.e.*, Eq. (2), to avoid excessive noise. A feasible way to solve this issue is that of modifying the 3DMM formulation so as to make the deformation components $\mathbf{D}$ sparse.

6 CONCLUSIONS

In this paper, we have proposed a 3DMM based solution to reconstruct a highly-detailed 3D model of the face starting from an RGB-D low-resolution face sequence. This is obtained by the combined effect of two specific algorithmic solutions for 3DMM construction and fitting: on the one hand, we used a dictionary learning based 3DMM implementation that makes possible modeling local deformations of the face; on the other, the model is fit to a target point cloud by a two steps approach, where the 3DMM is first deformed under the effect of the correspondence between a limited set of landmarks, and subsequently refined by an iterative local adjustment of point correspondences. Further, a robust denoising

solution is applied to the RGB-D sequences in order to preprocess sets of consecutive frames before applying the 3DMM fitting. Preliminary results have been reported that show the reconstruction errors between the 3D models derived after 3DMM fitting and the corresponding ground truth scans. It clearly emerges as the proposed framework provides superior results with respect to a landmark-based solution.

As further step in the direction of proposing our system for face rehabilitation purposes, we are collecting a face dataset that includes RGB-D sequences of the face captured by a Kinect camera, and the corresponding high-resolution scans acquired with a 3dMD scanner.

REFERENCES

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

Anasosalu, P. K., Thomas, D., and Sugimoto, A. (2013). Compact and accurate 3-D face modeling using an rgb-d camera: Let's open the door to 3-D video conference. In *IEEE Int. Conf. on Computer Vision Workshops*, pages 67–74.

Bas, A., Smith, W. A., Bolkart, T., and Wuhrer, S. (2016). Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision*, pages 377–391. Springer.

Berretti, S., Pala, P., and Del Bimbo, A. (2014). Face recognition by super-resolved 3D models from consumer depth cameras. *IEEE Trans. on Information Forensics and Security*, 9(9):1436–1449.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *ACM Conf. on Computer Graphics and Interactive Techniques*, pages 187–194.

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

Bondi, E., Pala, P., Berretti, S., and Del Bimbo, A. (2016). Reconstructing high-resolution face models from kinect depth sequences. *IEEE Trans. on Information Forensics and Security*, 11(12):2843–2853.

Booth, J., Roussos, A., Ververas, E., Antonakos, E., Ploumpis, S., Panagakis, Y., and Zafeiriou, S. (2018). 3D reconstruction of 'in-the-wild' faces in images and videos. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(11):2638 – 2652.

Booth, J., Roussos, A., Zafeiriou, S., Ponniahand, A., and Dunaway, D. (2016). A 3D morphable model learnt from 10,000 faces. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 5543–5552.

Chui, H. and Rangarajan, A. (2000). A feature registration framework using mixture models. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 190–197.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Ferrari, C., Lisanti, G., Berretti, S., and Del Bimbo, A. (2015). Dictionary learning based 3D morphable model construction for face recognition with varying expression and pose. In *Int. Conf. on 3D Vision*.

Ferrari, C., Lisanti, G., Berretti, S., and Del Bimbo, A. (2017). A dictionary learning-based 3D morphable shape model. *IEEE Trans. on Multimedia*, 19(12):2666–2679.

Hernandez, M., Choi, J., and Medioni, G. (2012). Laser scan quality 3-D face modeling using a low-cost depth camera. In *European Signal Processing Conf. (EUSIPCO)*, pages 1995–1999. IEEE.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. (2011). Kinectfusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *ACM Symposium on User Interface Software and Technology*, pages 559–568.

Kazemi, V., Keskin, C., Taylor, J., Kohli, P., and Izadi, S. (2014). Real-time face reconstruction from a single depth image. In *IEEE Int. Conf. on 3D Vision*, volume 1, pages 369–376.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009). Online dictionary learning for sparse coding. In *Int. Conf. on Machine Learning*, pages 689–696.

Myronenko, A. and Song, X. (2010). Point set registration: Coherent point drift. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A. (2011). Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE Int. Symposium on Mixed and Augmented Reality*.

Patel, A. and Smith, W. A. P. (2009). 3D morphable face models revisited. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1327–1334.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). 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.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. (2005). Overview of the face recognition grand challenge. In *IEEE Work. on Face Recognition Grand Challenge Experiments*, pages 947–954.

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

Zollhöfer, M., Martinek, M., Greiner, G., Stamminger, M., and Süßmuth, J. (2011). Automatic reconstruction of personalized avatars from 3D face scans. *Computer Animation and Virtual Worlds*, 22(2-3):195–202.

Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., and Stamminger, M. (2014). Real-time non-rigid reconstruction using an RGB-D camera. *ACM Trans. on Graphics*, 33(4):156:1–156:12.