Reconstructing High-Resolution Face Models From Kinect Depth Sequences

Enrico Bondi, Pietro Pala, *Senior Member, IEEE*, Stefano Berretti, *Member, IEEE*, and Alberto Del Bimbo, *Senior Member, IEEE*

***Abstract*— Performing face recognition across 3D scans with different resolution is now attracting an increasing interest thanks to the introduction of a new generation of depth cameras, capable of acquiring color/depth images over time. In fact, these devices acquire and provide depth data with much lower resolution compared with the 3D high-resolution scanners typically used for face recognition applications. If data are acquired without user cooperation, the problem is even more challenging, and the gap of resolution between probe and gallery scans can yield to a severe loss in terms of recognition accuracy. Based on these premises, we propose a method to build a higher resolution 3D face model from 3D data acquired by a low-resolution scanner. This face model is built using data acquired when a person passes in front of the scanner, without assuming any particular cooperation. The 3D data are registered and filtered by combining a model of the expected distribution of the acquisition error with a variant of the *lowess* method to remove outliers and build the final face model. The proposed approach is evaluated in terms of accuracy of face reconstruction and face recognition.**

***Index Terms*— Kinect depth camera, increased resolution, manifold estimation, locally weighted regression, face recognition.**

I. Introduction

PERSON identity recognition by the analysis of 3D face scans is attracting an increasing interest, with several challenging issues successfully investigated, such as 3D face recognition in the presence of non-neutral facial expressions, occlusions, and missing data [1], [2]. Existing solutions have been evaluated following well defined protocols on consolidated benchmark datasets, which provide a reasonable coverage of the many different traits of the human face, including variations of gender, age, ethnicity, and expressions, occlusions due to hair or external accessories, missing parts caused by pose changes. The resolution at which 3D face scans are acquired varies across different datasets, but given a dataset it is typically the same for all the scans. Due to this, the difficulties posed by considering 3D face scans with different resolutions and their impact on the recognition accuracy have not been explicitly addressed in the past. Nevertheless, there is an increasing interest for methods capable of performing recognition across scans acquired with different resolutions. This is mainly motivated by the availability of a new generation of low-cost, low-resolution 3D dynamic scanning devices (i.e., 3D plus time, also called 4D), such as Microsoft Kinect or Asus Xtion PRO LIVE. In fact, these devices are capable of a combined color-depth (RGB-D) acquisition at about 30fps, with an optimal working distance from the sensor ranging from $40cm$ up to $1.5m$. The spatial resolution of such devices is lower than that of high-resolution 3D scanners, but these latter are also costly, bulky and highly demanding for computational resources. Despite the lower resolution, the advantages in terms of cost and applicability of consumer cameras motivated some preliminary works performing face detection [3], re-identification [4], continuous authentication [5] and recognition [6]–[8] directly from the depth frames of the Kinect camera. However, based on the opposite characteristics evidenced by 4D low-resolution and 3D high-resolution scanners, new applicative scenarios can be devised, where high-resolution scans are likely to be part of gallery acquisitions, whereas probes are acquired with 4D cameras, resulting in lower resolution models.In this context, reconstructing a higher-resolution model out of a sequence of low-resolution depth frames is a plausible way to bridge the gap between low- and high-resolution acquisitions.

Based on these premises, in this work we define an approach that given a sequence of low-resolution depth frames reconstructs a higher-resolution face model. Some recent works explicitly addressed this problem [9], but require a cooperative protocol for the acquisition of the 3D dynamic sequence. Differently, in the proposed solution we aim to improve the previous work by removing the user cooperation requirement, and enabling the extraction of a 3D facial model of a person that just passes in front of the camera.

A. Related Work

The idea of constructing a higher-resolution representation of an object or scene from multiple low-resolution observations, possibly altered by noise, blurring or geometric warping, has been first introduced for 2D still images. Later, this concept has been extended to 3D generic data for recovering one high-resolution model from a set of low-resolution 3D acquisitions. For example, in [10] data acquired with a time-of-flight camera are upsampled and denoised by using information from a high-resolution image of the same scene taken from a viewpoint close to the depth sensor. Time-of-flight data are processed also in [11] by using an energy

Manuscript received November 17, 2015; revised April 30, 2016 and July 29, 2016; accepted August 3, 2016. Date of publication August 16, 2016; date of current version October 11, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Karthik Nandakumar.

The authors are with the Department of Information Engineering, University of Florence, 50139 Florence, Italy (e-mail: enrico.bondi@unifi.it; pietro.pala@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/TIFS.2016.2601059

minimization framework that explicitly takes into account the characteristics of the sensor, the agreement of the reconstruction with the aligned low resolution maps and a regularization term to cope with reconstruction of sparse data points. Some works on this topic also focus on 3D faces [12]–[14]. In [14], high-resolution 3D face models are used to learn the mapping between low- and high-resolution data. Given a new low-resolution face model the learned mapping is used to compute the high-resolution face model. Differently, in [13] the reconstruction process is modeled as a progressive resolution chain, whose features are computed as the solution to a maximum a posteriori estimation (MAP) problem. However, in both the cases, the framework is validated just on synthetic data. In [12], an algorithm is proposed that takes a single face frame from a Kinect depth camera, and produces a high-resolution 3D mesh of the input face. In this approach, the input depth frame is divided into semantically significant regions (eyes, nose, mouth, cheeks) and a database of high-resolution scans is searched for the best matching shape per region. The input depth frame is further combined with the matched database shapes into a single mesh that results in a high-resolution shape of the input person.

In the approaches above, the higher-resolution reconstruction depends on a single 3D low-resolution scan, with the additional information used for reconstruction coming from multiple high-resolution scans used as reference. This completely disregards the temporal dimension available in depth sequences acquired with a Kinect sensor. In order to exploit such temporal information, some methods approach the problem of noise reduction in depth data by fusing the observations of multiple scans [15]–[18]. In [17], the Kinect Fusion system is presented, which takes live depth data from a moving Kinect camera and creates a high-quality 3D model for a static scene. Later, in [19] dynamic interaction has been added to the system, where camera tracking is performed on a static background scene and a foreground object is tracked independently of camera tracking. Aligning all depth points to the complete scene from a large environment (e.g., a room) provides very accurate tracking of the camera pose and mapping [17]. However, this approach is targeted to generic objects in internal environments, rather than to faces. In [18], the approach is further extended to cope with non-rigid objects, such as faces, but results of non-rigid object denoising are demonstrated only in cases where the object to camera distance is almost constant. The work in [20] proposes to enhance low resolution dynamic depth videos containing non-rigidly moving objects with a dynamic multi-frame super-resolution algorithm. This is obtained by accounting for non-rigid displacements in 3D, in addition to 2D optical flow, and simultaneously correcting the depth measurement by Kalman filtering. This concept is incorporated in a multi-frame super-resolution framework, formulated in a recursive manner that ensures real-time deployment. Reported results range from a full moving human body to a dynamic facial video with varying expressions.In [16], a 3D face model with an improved quality is obtained by a user moving in front of a low resolution depth camera. The model is initialized with the first depth image, and then each subsequent cloud of 3D points is registered to the reference one using a GPU implementation of the ICP algorithm. This approach is used in [15] to investigate whether a system that uses reconstructed 3D face models performs better than a system that uses the individual raw depth frames considered for the reconstruction. To this end, authors present different 3D face recognition strategies in terms of the used probes and gallery. The reported analysis shows that the scenarios where a reconstructed 3D face model is compared against a gallery of reconstructed 3D face models, and where one frame (1F) is compared against multiple frames in the gallery, provide better results compared to the baseline 1F-1F approach. The approach proposed in [9] is the most closed solution to our method. Here the idea is to constructing a high-resolution face model from a sequence of low-resolution depth frames acquired with a Kinect camera. However, the approach leans strongly of the acquisition protocol, assuming the subjects sit in front of the camera at a predefined distance, moving the head to their left/right side in order to expose different parts of the face to the sensor. This avoids, a priori, scale and velocity problems, permitting a solution where the increased resolution of the reconstructed model can be obtained with up-sampling and 2D-Box splines approximation of the cumulated 3D point cloud obtained by rigid (ICP) registration of multiple 3D frames of a sequence.

B. Our Method and Contribution

In this paper, we present an original solution to derive one 3D face model from low-resolution depth frames acquired with a RGB-D sensor. In the proposed approach, first the face is automatically detected and cropped in each depth frame of a sequence, and the extracted 3D face data are aligned with each other so as to build a cumulated face model. Then, an initial denoising operation is performed, which is based on the anisotropic nature of the error distribution with respect to the viewing direction of the acquired frames. Finally, a manifold estimation approach based on the *lowess* non-parametric regression method is used to approximate the face surface from the cumulated face model and remove outliers from the data. The proposed approach has been evaluated on an extended version of the *Florence Superface* dataset [9], which includes depth sequences capturing the enrolled persons in cooperative as well as non-cooperative contexts, and high-resolution face scans acquired with a 3dMD scanner.

In summary, the main contributions of this paper are:

- An approach to reconstruct a 3D face model from a sequence of low-resolution depth frames of the face that can work both in cooperative and non-cooperative contexts, with the only constraint of having the face of the subjects in the operating field of the camera. After entering the field of view of the camera, the user can get closer to it and change the orientation of the face with respect to the camera, yet maintaining the face exposed to it. This requires specific solutions to manage pose and velocity variations in a sequence. The resolution of the reconstructed face model is higher than the resolution of the individual depth frames;
- A thorough evaluation demonstrating the accuracy of the reconstructed face models. This is quantitatively

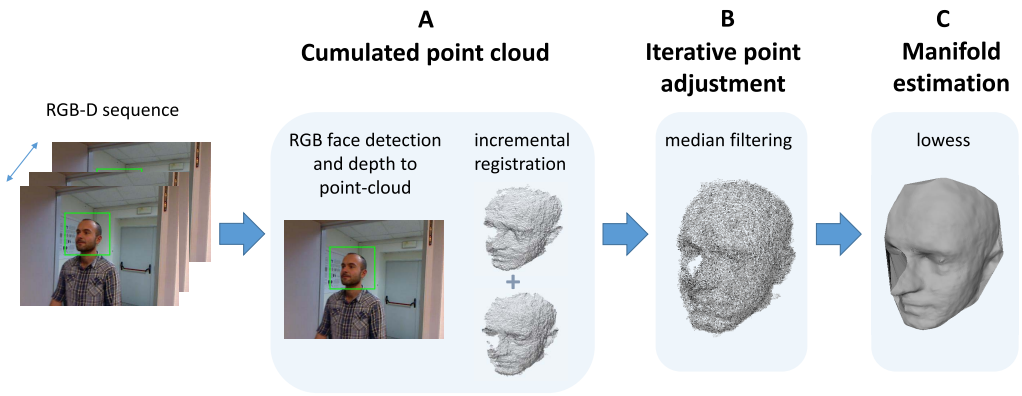


Fig. 1. Processing steps applied to the RGB-D sequences. In (A), the face is first detected in each RGB frame, and the corresponding data in the depth frame are transformed to a point cloud; The point cloud are then registered together so as to obtain a *cumulated point cloud* capturing all the data of a sequence; In (B), the 3D positions of the points of the cumulated point cloud are adjusted using an iterative procedure based on median filtering; In (C), a 2D-manifold of the face is estimated (reconstructed) by applying a *local weighted regression* to the principal components of the cumulated point cloud.

measured both in terms of mean distance error between the reconstructed and the high-resolution facial surface and in terms of recognition rate against a gallery of high-resolution scans.

The proposed method develops on our previous studies on reconstruction of super-resolved face models in a cooperative scenario [9]. In particular, we introduce a novel face reconstruction method that relaxes the constraint of the cooperative scenario: During data acquisition, the user is not requested to move the head following some predefined motion pattern; his/her distance to the camera may change; and even the part of the face that is exposed to the camera can vary. The improvements of the proposed solution are threefold: *(i)* registration of point clouds is achieved by combining the Iterative Closest Point [21] and the Coherent Point Drift [22] approaches so as to cope with slight scale variations (due to the approximations of the sensor calibration) and face deformations; *(ii)* the registered point cloud is subject to noise removal through a nonlinear filtering scheme, which exploits the anisotropic distribution of the acquisition error; *(iii)* local weighted regression is used in place of global box-spline approximation to reconstruct the face surface. This scheme yields a better fit to local variations of the surface compared to box-splines used in [9]. Preliminary ideas and results related to the proposed method were first reported in [23]. With respect to that previous work, the manifold reconstruction problem is now addressed by combining lowess estimation with a median filtering, which exploits the anisotropic error distribution of the range sensor. This latter one is a completely new contribution of this paper that allows us to properly initialize the lowess iterative procedure resulting in a more effective and efficient solution.In addition, we revised and extended the experimental evaluation of the proposed approach, which is demonstrated using more people observed from multiple viewpoints, and with application to recognition and re-identification tasks. A comparative analysis with [9] and [17] is also presented.

The rest of the paper is organized as follows: The scenario and the problem statement are defined in Sect. II; Sect. III reports the overall point cloud construction by registering the depth frames of a sequence, and the initial removal of noisy points based on anisotropic error modeling; Face reconstruction based on manifold estimation is presented in Sect. IV; Experimental results are reported and discussed in Sect. V, focusing on a realistic scenario of data acquired at an access gate; Finally, conclusions are drawn in Sect. VI.

II. Problem Overview

In this work, we aim at reconstructing a *higher-resolution* 3D model of the face, by processing a sequence of low-resolution *depth frames* (*frames* in the following) acquired with a Kinect camera. The depth sequences acquired with the RGB-D camera feed the processing pipeline sketched in Fig. 1.

In the initial step, the face is detected in each RGB frame of the input sequence using a state of the art face detector [24]. The face detector is capable of correctly detecting faces in frontal and side views up to 90°, over a broad range of scales spanning from less than one to a few meters. By exploiting the fact that the RGB and depth frames are registered with each other, the face region detected in the RGB frame is projected in the depth frame, so as to extract the depth data about the detected face and convert them into a point cloud in the 3D (X, Y, Z) coordinate system of the camera. The point clouds extracted from different depth frames are registered with each other, thus obtaining a *cumulated point cloud*, which includes all the depth data of the sequence (block *A* in Fig. 1). The cumulated point cloud is then processed to reduce noise and construct the higher resolution face model, passing through two additional steps, namely, *iterative point adjustment* along the line of sight of the camera, and *local weighted regression* for manifold estimation (corresponding, respectively, to the blocks *B* and *C* in Fig. 1). The iterative point adjustment relies on the characteristics of the RGB-D camera, by which the (x, y, z) coordinates of a generic point of the cloud are affected by an acquisition error, which is anisotropically distributed (i.e., the variance of the error along the Z axis, aligned to the line of sight, is much larger than the variance along the X and Y directions) [25], [26]. This suggested us it is possible to adjust the position of the points in the cumulated point cloud by exploiting their multiple acquisitions from different viewing directions. The final step

estimates the 2D-manifold of the face, by regularizing the cumulated point cloud based on the non-parametric local weighted regression (*lowess*) method.

Details on the steps above are given in the following.

III. Cumulated Point Cloud

The first step to compute a high-resolution model of the face is to cumulate the data of individual frames in a registered way, thus deriving a denser point cloud (compared to the density of data in a single frame) from multiple observations of the face. Then, the dense point cloud is processed to discard redundant or noisy observations.

A. Incremental Registration

Let $k \in \{1, \ldots, K\}$ be the indexes of the frames where a face is detected, and $\mathbf{p}_i^{(k)}$ the 3D coordinates (x, y and the depth value z) of the i-th observed facial point in the k-th frame. Registration of the 3D facial data is operated starting from the first frame where a face is detected. It should be noticed that using the 3D coordinates, which express the position of points in real coordinates (millimeters) instead of image coordinates, makes the measurements independent on scale. The first order cumulated point cloud $\mathcal{C}^{(1)}$ merges facial data extracted from the first two frames:

$$\mathcal{C}^{(1)} = \mathcal{R}^{(1)}\left(\{\mathbf{p}_i^{(1)}\}_i, \{\mathbf{p}_i^{(2)}\}_i\right) \bigcup \{\mathbf{p}_i^{(1)}\}_i, \tag{1}$$

being $\mathcal{R}(S_1, S_2)$ the registration operator that returns the points in the second set S_2 after registering them to the points in S_1. This registration operator should cope with quite general depth sequences, allowing the users to walk and/or change the pose of the head, so that the scale and pose of the face can vary during acquisition. To account for these difficulties, the registration operator is obtained by cascading the Coherent Point Drift (CPD) algorithm [22], a probabilistic method for non-rigid registration of point sets, and the Iterative Closest Point (ICP) algorithm [21], which performs rigid registration between point sets. This combination resulted in the best alignment between subsequent point clouds. In fact, CPD is demonstrated to be more robust to noise and outliers than standard rigid registration methods [22]. In addition, CPD can cope with non-rigid deformations, such as those related to facial expressions and/or speaking that can occur in an unconstrained acquisition process. So, running CPD permitted us to obtain a first good initialization of the alignment, which is subsequently refined with the rigid ICP registration.

Using this procedure, data in the next frame $\{\mathbf{p}_i^{(3)}\}_i$ are aligned to the first order cumulated point cloud to yield the second order cumulated point cloud:

$$\mathcal{C}^{(2)} = \mathcal{R}^{(2)}\left(\mathcal{C}^{(1)}, \{\mathbf{p}_i^{(3)}\}_i\right) \bigcup \mathcal{C}^{(1)}. \tag{2}$$

This registration process is iterated for all the available frames, yielding the K-th order cumulated point cloud $\mathcal{C}^{(K)}$.

Figure 2 shows the effect of cumulating point clouds of subsequent frames. In (a), the point cloud of a frame is reported in red, whereas in (b) the current cumulated point cloud (constructed with the frames preceding the frame in (a)) is reported; in (c) the final result is shown, after the point cloud in (a) is registered and added to the cumulated point cloud of the sequence using CPD and ICP. It should be noticed that the construction of a cumulated point cloud allows the proposed method to cope with large pose variations with respect to the first observed frame. In fact, the presence of a large pose variation between the first and current observation is compensated by the combination of all the observations between the first and the current one.

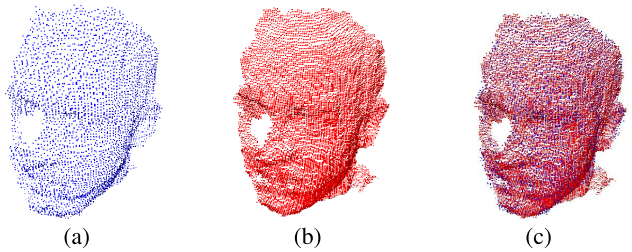


Fig. 2. Cumulated point cloud construction: (a) Point cloud of a frame in blue; (b) Cumulated point cloud of the sequence; (c) Final result obtained by registering and cumulating the point cloud in (a) to the point cloud in (b) using CPD and ICP.

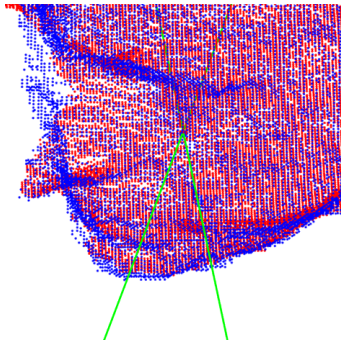


Fig. 3. Two point clouds (one with red and one with blue colors) and the directions of the lines of sight of two points.

B. Iterative Point Adjustment

The cumulated point cloud is composed of points extracted from different frames of the acquisition sequence, and thus observed from different viewing directions—this corresponds to the most general case of a subject that moves with respect to the camera.

As an example, Fig. 3 shows the cumulated point cloud obtained from two different frames (red and blue colors are used based on the frame points are extracted from), and the lines of sight of two points of the cloud—the two points are actually two observations from different directions of the same point on the face surface. Given a point of the cloud, its line of sight identifies the direction along which the maximum expected measurement error can be observed [26]. Thus, regularization of the point cloud is accomplished by moving each point along its line of sight, so as to maximize the consistency between the coordinates of the point and its neighbors.Formally, given a generic *pivot* point of the cumulated cloud $\hat{\mathbf{p}} \in \mathcal{C}^{(K)}$, its estimated true position $\mathbf{p}_e$ is measured as the median value of points of the cloud that lie inside the cylinder $C_r(\hat{\mathbf{p}})$ of radius r and axis aligned with the

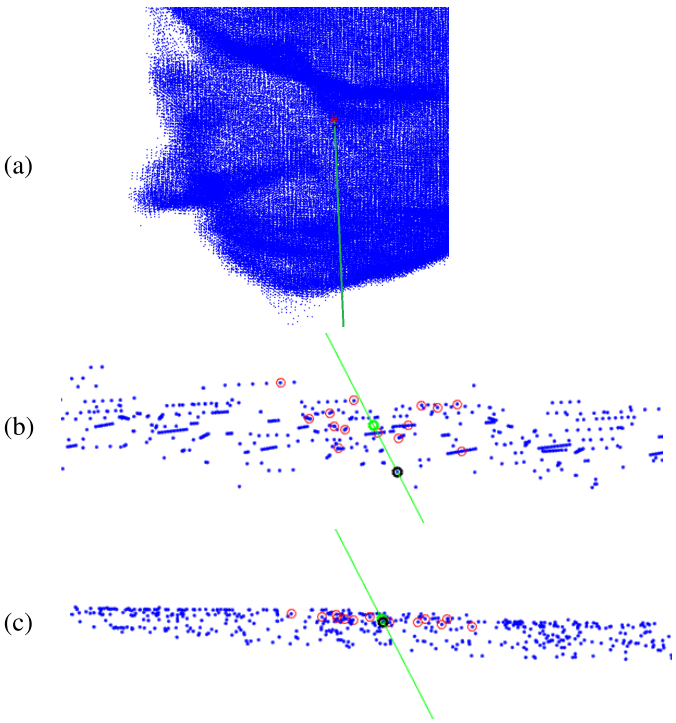


Fig. 4. (a) The cumulated point cloud and points of the cloud (red circles) that are within the cylinder aligned with the line of sight of the pivot point (black circle); (b) Close up and top view of the cylinder in (a) with highlighted the line of sight of the pivot (green line), the points of the cloud within the pivot cylinder (red circles), and the projection of the median of points of the cloud within the pivot cylinder onto the line of sight (green circle); (c) Close up and top view of data shown in (a) after one adjustment step.

line of sight of $\hat{\mathbf{p}}$:

$$\mathbf{p}_e = median\left\{\mathbf{p} \in \mathcal{C}^{(K)} \bigcap C_r(\hat{p})\right\}. \tag{3}$$

By using the median operator to estimate the true coordinates of the pivot, the influence of outliers is reduced. Filtering of the point cloud is operated through an iterative procedure that adjusts the position of each point of the cloud toward its estimated true position. At each iteration, all points of the cloud are processed and their positions adjusted. To reduce the computational cost of the iterative procedure, points of the cloud are organized in a kd-tree index structure. Convergence has been obtained, on average, with two iterations. The stopping criteria uses a threshold on the variation between two subsequent iterations.

As an example, Fig. 4 shows a general view and two close up views of points of the cloud within the cylinder computed for one of the point of the cloud, acting as the pivot (a value of r=1.5mm for the radius of the cylinder has been used). The two close up views show that after just one adjustment step of all the points of the cloud (Fig. 4(c)), approximation of the thin face surface is more accurate than in the original cloud (Fig. 4(b)). This is particularly true if points within the pivot cylinder are considered (red circles).

IV. Manifold Estimation

The result of the registration and adjustment processes described in the previous Sections is a point cloud that collects a set of points in the 3D space. The generic i-th point $\mathbf{p}_i = (x_i, y_i, z_i)$ can be regarded as the observation, affected by some noise, of the underlying face surface that can be modeled as a 2D-manifold embedded in the 3D space. In the proposed approach, reconstruction of the true face surface is formalized as a problem of manifold estimation from noisy data. For this purpose, we adopt an approach based on the combination of *dimensionality reduction* and *local weighted regression*, similarly to [27] and [28] (see also block *C* in Fig. 1). Mapping 3D data into the 2D embedding makes it explicit the distance on the manifold rather than in the 3D space. This is used to compute, for a generic point of the cloud, the set of its closest neighbors based on the local geometry of the manifold.

A. Dimensionality Reduction

Principal Component Analysis (PCA) [29] is used to reduce the dimensionality of the manifold and compute a 2D-embedding of the 3D point cloud. In this way, the intrinsic geometry of the manifold is preserved by mapping close points on the manifold (that does not necessarily mean close points in the 3D space) into close points on the 2D embedding. In general, other dimensionality reduction methods could be combined with the proposed framework. We also considered *Isomap* [30] as a candidate solution, since it preserves in the embedded space the geodesic distances between points on the manifold. However, comparison of the results obtained with Isomap and PCA did not show considerable differences, and the much lower computational complexity of PCA induced us to adopt this latter approach.

More in detail, being M the number of points in the cumulated point cloud, their average is computed:

$$\bar{\mathbf{p}} = \sum_{i=1}^{M} \mathbf{p}_i, \tag{4}$$

and subtracted from the observations:

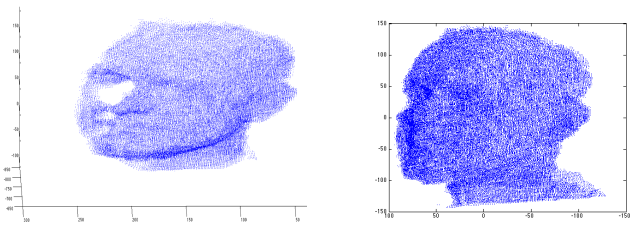
$$\mathbf{p}'_i = \mathbf{p}_i - \bar{\mathbf{p}}, \quad i = 1, \ldots, M. \tag{5}$$

A matrix $\mathbf{P} \in \mathbb{R}^{3\times M}$ is then constructed, with the points $\mathbf{p}'_i$ as columns.Performing PCA of the covariance matrix $\mathbf{C} = \mathbf{P}\cdot\mathbf{P}^t$, the matrix $\mathbf{U} \in \mathbb{R}^{3\times 3}$ is determined, whose columns are the eigenvectors of $\mathbf{C}$. Then, the two eigenvectors corresponding to the two largest eigenvalues of $\mathbf{C}$ are considered as columns of the matrix $\mathbf{U}_2$, which spans the 2D embedding subspace. Finally, the projection of the cumulated point cloud in the embedding subspace can be computed as $\mathbf{Q} = \mathbf{U}_2^t \cdot \mathbf{P} \in \mathbb{R}^{2\times M}$.

As an example, Fig. 5(a)-(b) show, respectively, the 3D points of the cumulated point cloud, and the corresponding 2D-embedding using PCA.

B. Locally Weighted Regression

Let $\mathbf{q}_i = (u_i, v_i) \in \mathbb{R}^2$ be the coordinates of the 3D points $\mathbf{p}_i$ after projection onto the 2D-embedding. Following the original approach described by Cleveland [31], estimation of the manifold at point $\mathbf{p}_i$ is accomplished by fitting a low-dimensional polynomial to the subset of points of the cumulated point cloud that are mapped close to $\mathbf{q}_i$ on the 2D-embedding.Operatively, the subset of data is determined



(a) (b)

Fig. 5. (a) Cumulated point cloud in 3D; (b) 2D-embedding of the points in (a) using PCA.

by a nearest neighbors algorithm on the 2D-embedding. The cardinality of this subset is controlled through a *smoothing parameter* $\alpha \in (0, 1)$. The points used to fit the polynomial are the αM closest to $\mathbf{q}_i$ on the 2D-embedding. This set is denoted as $N(\mathbf{q}_i)$. Large values of α produce smooth regression functions that wiggle the least in response to fluctuations in the data. The smaller α is, the closer the regression function will conform to the data, thus yielding poor robustness to noise.

For each point $\mathbf{q}_i$ $(i = 1, \ldots, M)$, a weight $w_j(\mathbf{q}_i)$ is computed using a *tricube* weight function:

$$w_j(\mathbf{q}_i) = \begin{cases} \left(1 - h_j^3\right)^3 & if \quad h_j \in (0, 1) \\ 0 & otherwise, \end{cases} \tag{6}$$

where h_j is the distance between $\mathbf{q}_i$ and $\mathbf{q}_j$ in the 2D embedding, scaled by the maximum distance between points $\mathbf{q}_j \in N(\mathbf{q}_i)$ and $\mathbf{q}_i$:

$$h_j = \frac{d\left(\mathbf{q}_i - \mathbf{q}_j\right)}{\max_{\mathbf{q}_j \in N(\mathbf{q}_i)} d\left(\mathbf{q}_i - \mathbf{q}_j\right)}. \tag{7}$$

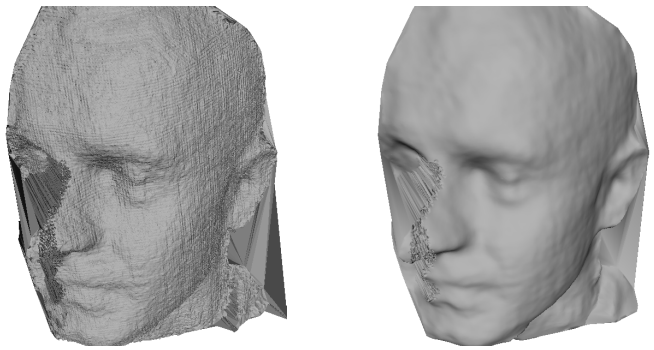
It should be noticed that Eq. (6) gives higher weight values to points that are close to $\mathbf{q}_i$, and zero value to weights of the points that are outside $N(\mathbf{q}_i)$. Furthermore, based on Eq. (6) the value of the weight of the central point is zero $(w_i(\mathbf{q}_i) = 0)$. In this way, the regression function used to project $\mathbf{q}_i$ onto the 3D space depends only on the points in the neighbor $N(\mathbf{q}_i)$ and not on $\mathbf{q}_i$ itself. This choice is intended to increase robustness to outliers.

These weights are used to approximate the function that locally maps points of the 2D embedding (i.e., $\mathbf{q}_j$) onto the manifold (i.e., $\mathbf{p}_j$) through a local weighted regression scheme. This is obtained by fitting a second order polynomial on pairs $\mathbf{q}_j \mapsto \mathbf{p}_j$, weighted by $w_j(\mathbf{q}_i)$, for each $\mathbf{q}_j \in N(\mathbf{q}_i)$.

Algorithm 1 Locally Weighted Regression Fit

Require: $\langle \alpha, M, \mathbf{q}_i \rangle$
Ensure: $\hat{\mathbf{p}}_i$
for all $\mathbf{q_i}$, $i = 1, \ldots, M$ **do**
compute set $N(\mathbf{q}_i)$
compute weights $w_j(\mathbf{q}_i)$ for points in $N(\mathbf{q}_i)$
solve the weighted least squares to identify the polynomial $\mathcal{P} : \mathbb{R}^2 \mapsto \mathbb{R}^3$
update the coordinates of the i-th point of the cloud $\hat{\mathbf{p}}_i = \mathcal{P}(\mathbf{q}_i)$
end for

Algorithm 1 summarizes the estimation process. In particular, the smoothing parameter α has been set equal to 0.0035,



(a) (b)

Fig. 6. Effect of Locally Weighted Regression filtering for surface reconstruction: (a) the face surface is reconstructed using only the Iterative Point Adjustment; (b) the face is reconstructed applying Locally Weighted Regression after Iterative Point Adjustment.

TABLE I

ESTIMATED AND MEASURED COMPUTATIONAL COMPLEXITY OF THE PROPOSED RECONSTRUCTION METHOD (TIME MEASURES ARE REFERRED TO A MATLAB CODE RUNNING ON A 3.2GHz CPU WITH 32Gb MEMORY). K IS THE FRAME IMAGE SIZE; S AND C ARE THE NUMBER OF POINTS IN THE CURRENT FRAME AND CUMULATED POINT CLOUD, RESPECTIVELY; M IS THE NUMBER OF POINTS IN THE OVERALL POINT CLOUD

step	O(.)	$time(s)$
face detection	$O(k)$	0.067
frame registration (CPD+ICP)	$O(S * C)$	3.2
iterative point adjustment	$O(M^2 \log M)$	81.0
PCA	$O(M)$	0.04
lowess filtering	$O(M^2 \log M)$	130.0

while the number of points in the reconstructed model M resulted, on average, equal to 90000. The number of points in the neighbor $N(\mathbf{q}_i)$ of each point $\mathbf{q}_i$ is thus of about 300 points (i.e., αM).

Figure 6(b) shows the application of the locally weighted regression module to surface reconstruction. Compared to the reconstruction based only on iterative point adjustment, shown in Fig. 6(a), the surface is smoothed, yet preserving the details of the 3D geometry characterizing facial traits, such as the shape of the mouth, the nasal, the orbital and the auricular regions.

The computational complexity of the proposed reconstruction method has been estimated in Table I (some of the steps reported are repeated for individual frames of the sequence– on average, models are reconstructed from 23 frames). Currently, the processing is not real-time, the most onerous steps being the *iterative point adjustment* and *filtering*. However, these steps operate locally on the point cloud and could be parallelized.

V. EXPERIMENTAL RESULTS

To evaluate the reconstruction accuracy of the proposed approach, two different aspects have been considered: the *metric accuracy*, which measures the error of the reconstructed face model with respect to the high-resolution face model of

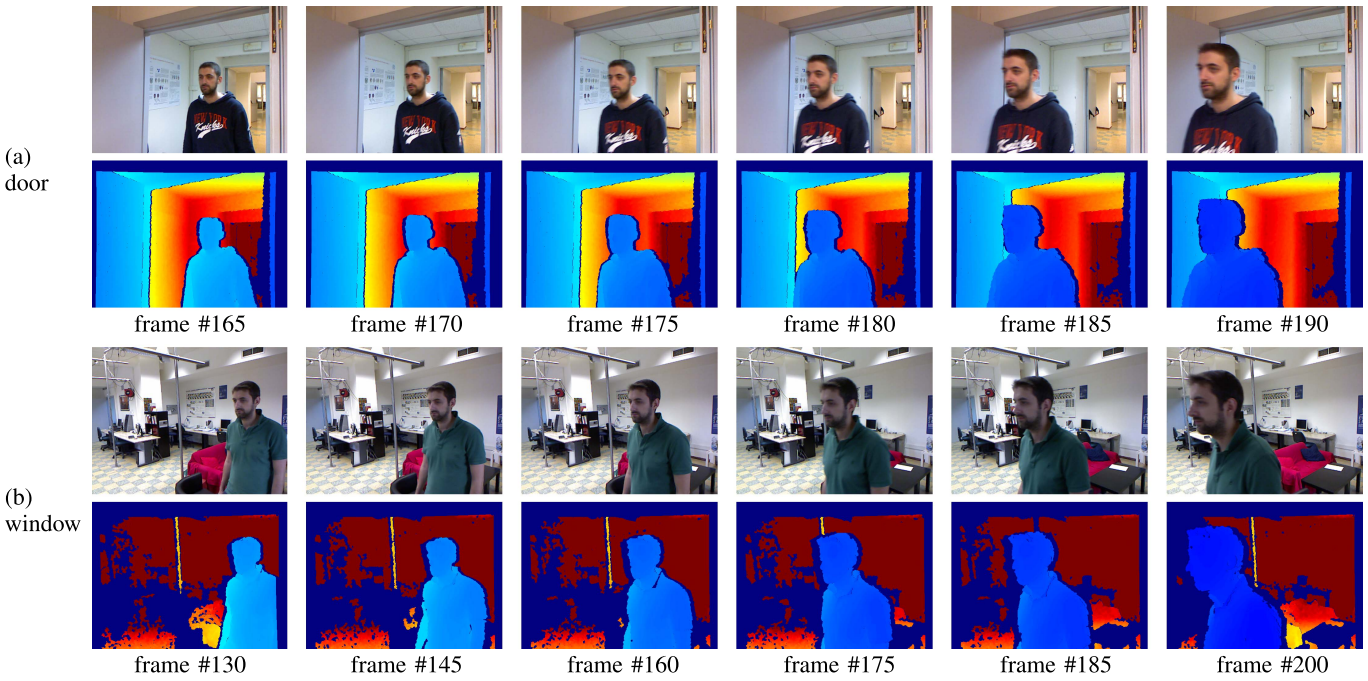


Fig. 7. Un-cooperative acquisition. Sample RGB and depth frames (false colors) of a same subject acquired with the Kinect camera in two different set-up: (a) *door*; (b) *window*.

the same person (Sect. V-B); the *recognition accuracy*, which measures the improved accuracy of face recognition using as probe a reconstructed face model instead of one or multiple low-resolution frames (Sect. V-C). The recognition accuracy is measured considering two distinct experiments: *identification*, and *re-identification*.

The *Florence Superface* dataset has been used for the evaluation. As a contribution of this work, this dataset has been extended to include new acquisitions captured according to an un-cooperative protocol, as described in the following.

A. The Florence Superface Dataset

Some public datasets exist for face analysis from consumer depth cameras like Kinect. Examples are the *EURECOM Kinect Face* dataset [32], or the *The 3D Mask Attack* database specifically targeted to detect face spoofing attacks [33]. However, the former only includes cooperative acquisitions, where the subjects stay in front of the camera at a predefined (fixed) distance, while the latter yet includes cooperative acquisitions only, and is devised to investigate mask attacks of face recognition methods based on Kinect data.Furthermore, since these datasets do not provide high-resolution scans of the enrolled subjects—just low-res data—they cannot be used to evaluate the accuracy of the proposed solution. Due to this, in the experiments reported hereafter, we decided to use and extend the *Florence Superface* dataset (UF-S) [9]. This dataset was originally designed to include 3D high resolution face scans, and 2D videos of the face acquired in different conditions [34]. Successive extensions of the dataset addressed the inclusion of depth video sequences of the face, acquired with the Kinect camera according to a cooperative protocol [9], [35]. In this work, we further extend this dataset by capturing depth video sequences for a subset of the subjects according to an un-cooperative protocol. In particular, the part of the UF-S used in the experiments includes the following data for each one of 25 different persons:

- A 3D high-resolution scan of the face of the person, with about 40,000 vertices acquired with a 3dMD scanner (see Fig. 8(c) for some examples). The geometry of the mesh is highly accurate with an average RMS error of about 0.2mm;
- Two un-cooperative *Kinect* video sequences (RGB-D), acquired in two slightly different conditions (called in the following *door* and *window*), where the person goes through an access point monitored by the RGB-D camera. The camera is mounted on a doorjamb or an easel, in the *door* and *window* set-up, respectively, at a height of about 170*cm*, well positioned for viewing the face of a person walking through the access (see Fig. 7(a)-(b)). The face is almost completely visible at the maximum working distance (about 180*cm*), while just a side-part of the face is exposed to the sensor when the person gets closer to the camera (minimum distance of about 40*cm*). It should be noticed that the size of the face changes as the person moves toward the camera, requiring the approach to cope with scale variations in the reconstruction process.

B. Metric Accuracy

This experiment aims to evaluate the error of the reconstructed 3D model with respect to the 3D high-resolution scan of the same subject. To better understand the accuracy of reconstruction, this error is compared to the error between the

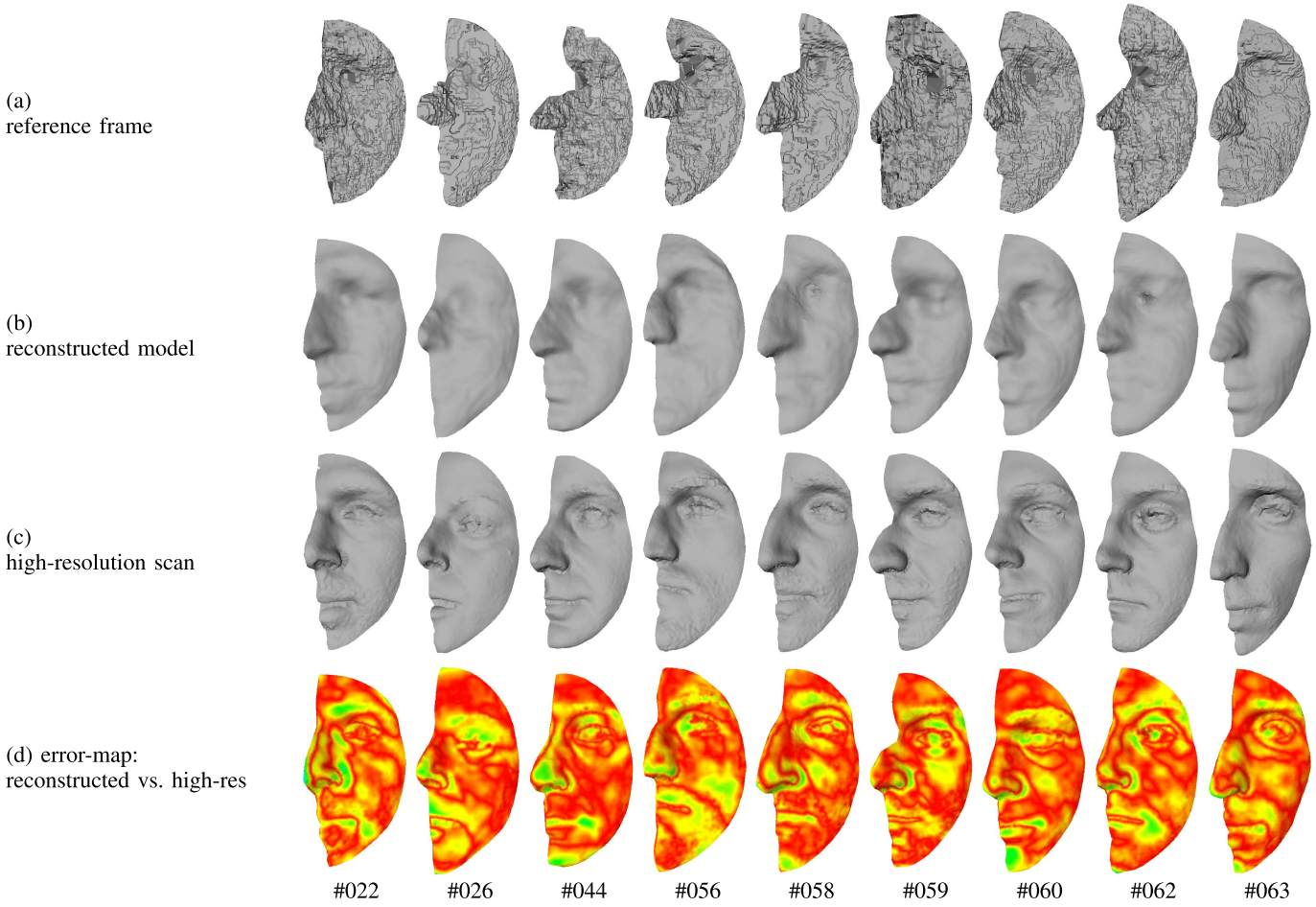


Fig. 8. For each subject in a column, we report: (a) The low resolution 3D scan of the reference frame; (b) The reconstructed 3D model; (c) The high-resolution 3D scan; (d) The error-map showing, for each point of the reconstructed model, the value of the distance to its closest point on the high-resolution scan after alignment (distance increases from red/yellow to green/blue).

depth data extracted from a *reference frame* of the sequence and the 3D high-resolution scan. For this purpose, the last frame where the face is detected in the RGB data is used as a reference. This case corresponds to the *best* condition for depth frame acquisition, where the subject is most close to the sensor, thus yielding the highest resolution for 3D facial data in a single depth frame.

For each subject used in the experiments, we considered: The high-resolution scan; The reconstructed model using the proposed approach; and the low-resolution scan obtained from the reference frame. In all these cases, the 3D facial data are represented as a mesh and cropped using a sphere of radius 95mm centered at the nose tip (the approach in [36] is used to detect the nose tip).Furthermore, to fully simulate the most general case of uncooperative data acquisition, we assume that just one half of the face is exposed to the sensor.Accordingly, we estimate the plane passing from the nose tip and ridge that divides the face into its right and left parts. This plane is used to retain 3D data about the part of the face that is most visible. To measure the error between the high-resolution scan and the reconstructed model of the same subject, they are first aligned through CPD registration [22]. Then, for each point of the reconstructed model its distance to the closest point in the high-resolution scan is computed to build an error-map. As an example, Fig. 8 shows the cropped 3D mesh of the reference frame, the reconstructed model, the high-resolution scan and the error-map between the reconstructed model and the high-resolution scan, for some test subjects (one column per subject).

To represent the average error of the reconstructed models and reference frames with respect to the high-resolution scans, the *Root Mean Square Error* (RMSE) between their surfaces S and S' is computed considering the vertex correspondences defined by the CPD registration, which associates each vertex $p \in S$ to the closest vertex $p' \in S'$:

$$RMSE(S, S') = \left(\frac{1}{N} \sum_{i=1}^{N} (p_i - p_i')^2 \right)^{1/2}, \tag{8}$$

being N the number of corresponding vertices in S and S'.

Results obtained using this distance measure are summarized in Table II. In particular, we reported the average values for the RMSE computed between the high-resolution scan and, respectively, the reconstructed model and the reference frame. On the one hand, values in Table II measure the magnitude of the error between the reconstructed model and the high-resolution scan of the same subject; On the other hand, they give a quantitative evidence of the increased quality of the

TABLE II
STATISTICS OF THE RMSE COMPUTED BETWEEN REFERENCE FRAMES AND RECONSTRUCTED MODELS WITH RESPECT TO 3D HIGH-RESOLUTION SCANS

comparison	RMSE min	max	mean	std dev
reference frame vs. high-resolution	0.96	2.02	1.51	0.28
reconstructed vs. high-resolution	0.79	1.21	1.11	0.14

TABLE III
RMSE COMPUTED BETWEEN REFERENCE FRAMES (*ref.*) AND 3D MODELS RECONSTRUCTED (*rec.*) WITH DIFFERENT SOLUTIONS WITH RESPECT TO 3D HIGH-RESOLUTION SCANS (*high-res*)

comparison	RMSE min	max	mean	std dev
ref. (first) frame vs. high-res	0.96	2.02	1.51	0.28
rec. Kinect fusion [17] vs. high-res	1.08	1.76	1.41	0.21
rec. super-res [9] vs. high-res	0.80	1.36	0.99	0.15
rec. median vs. high-res	0.82	1.42	1.03	0.15
***rec. lowess* vs. high-res**	**0.79**	**1.21**	**0.95**	**0.11**

reconstructed model with respect to the reference scan. This latter result is indeed an expected achievement of the proposed approach, since the reconstructed models combine information of several frames of a sequence. Thanks to the proposed processing pipeline, the mean error of the reconstructed model is considerably lower (less than 36% on average) than the mean error observed in the reference frame.

1) Comparative Evaluation: The metric results obtained for our approach have been also compared with a variant of our method that uses *median* filtering instead of *lowess* for noise removal, and two alternative state of the art solutions, namely, the *face Super-resolution* approach [9], and the *Kinect-fusion* approach proposed in [17].

Since the approaches in [9] and [17], are targeted mostly for contexts where the framed subject is slowly moving, we acquired a new set of depth videos with less uncooperative conditions, so as to enable these methods to work properly (i.e., the subjects move towards the camera keeping a frontal pose). Results of the comparison are reported in Table III. It can be observed that, on average, results of the proposed method using *lowess* outperform other solutions. In particular, *lowess* results to be more effective in removing noise, while capturing the local manifold of the reconstructed surface. For what concerns the other methods, the *super-resolution* approach in [9] ranks as the second best, with the *Kinect fusion* solution improving not so much with respect to rough frames.

Examples of reconstructed models obtained using the approaches listed above are also reported in Fig. 9.

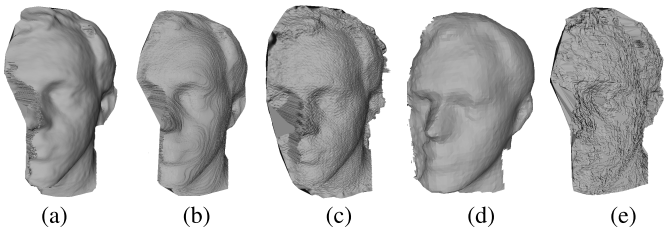


Fig. 9. Example models reconstructed using, respectively: (a) the proposed approach with *lowess* filtering; (b) the proposed approach with *median* filtering; (c) the *super-resolution* method in [9]; (d) the *Kinect fusion* solution [17]. In (e), the first frame of the sequence (i.e., *reference frame*) is shown.

C. Recognition Accuracy

The reconstructed 3D face better represents the shape of the face and is thus expected to enable more robust recognition compared to the use of 3D data extracted from a single frame. The potential of the proposed approach to enable accurate recognition is investigated with respect to two distinct scenarios: identification and re-identification.

In the subject identification task, the gallery is composed of high-resolution scans, whereas reconstructed models obtained in both the *door* and *window* set-up are used as probes (25 subjects, 2 models per subjects, one for the door and one for window set-up). Description and matching of gallery and probe models is obtained according to the face recognition approach proposed in [37], that is based on the extraction and comparison of local features of the face.

We included 66 high-resolution scans in the gallery, and considered the reconstructed models as probes (50 probe models in total). We remark here that the probe models reconstruct one side of the face thus making the recognition more difficult with respect to the case of full frontal faces. The recognition accuracy is evaluated through the Cumulative Matching Characteristic (CMC) curves. Figure 10 reports the CMC curve in the cases the reference frames (baseline) or the reconstructed models are used as probes. The blue curve clearly shows that using the reconstructed models a much higher recognition accuracy is achieved compared to the use of raw frames (red curve): the rank-1 recognition rate increases from about 32% to 78% and rank-10 from 74% to 96%.

To further motivate the advantage of reconstructing a model with increased resolution with respect to using raw depth frames, we also performed a recognition experiment where a set of N low-resolution depth frames is considered as representing a unique probe identity, and each gallery identity is instead represented by a high-resolution scan. In this way, the match between a probe and a gallery identity is regarded as a match of a probe sequence against a gallery scan (i.e., N vs. 1 match). A fusion mechanism based on the sum of ranking is used to produce the *probe* vs. *gallery* final ranking (we also tried *voting*, as fusion method, but the sum of rank provided better results).

In the subject *re-identification* task described in the following, both the gallery and the probes are composed of reconstructed models. Description and matching of gallery and probe models is obtained adopting the same approach used for the identification experiment (i.e., the method in [37]). For each one of the 25 different subjects, two different reconstructed models and two different low resolution reference frames are available, acquired in the *door* and *window* set-up, respectively. Half of the models/frames are randomly selected

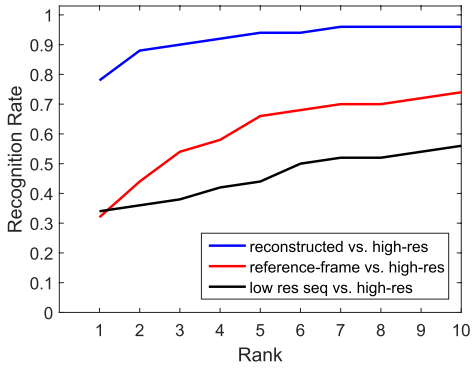


Fig. 10. CMC curves obtained by using high-resolution scans as gallery, and reconstructed models (blue curve), reference frames (red curve), and sequences of low resolution frames (black curve) as probes.

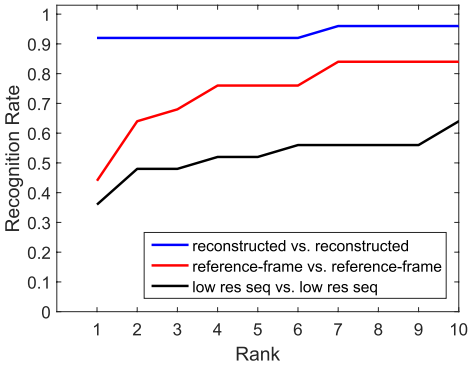


Fig. 11. CMC curves for the *re-identification* experiment. The red and the black curves represent the baselines, respectively, for the case of matching between individual low-resolution face frames (reference frames), and for the case in which full sequences of low resolution frames are used for both gallery and probes. The blue curve refers to the proposed approach, where reconstructed face models are used for both gallery and probes.

and used as gallery, while the remaining models/frames are used as probes. The baseline is evaluated using two methods: matching between two reference frames, one used as probe and one as gallery; and matching full sequences of low resolution frames. In the latter case, a set of N low-resolution depth frames is considered as representing a unique probe, and each gallery identity also comprises a set of M low-resolution depth frames. In this way, the match between a probe and a gallery identity is regarded as a match of two sequences, which is ultimately handled as a match between N vs. M frames. The sum of ranking is used as fusion mechanism.Results obtained with the baseline methods are compared with those obtained with the reconstructed models in Fig. 11 using CMC curves. From these curves, it clearly emerges that the proposed face reconstruction approach enables much higher face re-identification accuracy than the baseline solutions.

In both the recognition and re-identification scenarios, reported in Fig. 10 and Fig. 11, it can be noticed that matching more low-resolution frames of a sequence produces similar results to the case where only the first frame (reference) of the sequence is used. This is mainly motivated by the fact that, in the sequences, subjects enter the field of view of the camera at a distant point, go closer to the camera, and then move away from the camera. So, there is a sort of tradeoff between the face details that can be captured by Kinect (closer frames correspond to higher details) and the portion of the face which is visible to the sensor (closer frames include just a partial, side view of the face). Results suggest that performing face matching with just a part of the face frame adds noise to the results, yielding no improvement in terms of accuracy.

VI. Conclusions

In this paper, we have defined an approach that permits the construction of a higher-resolution face model starting from a sequence of low-resolution 3D scans acquired with a consumer depth camera in an uncooperative scenario. In the proposed framework, first the low-resolution 3D frames of a sequence are aligned using the Coherent Point Drift (CPD) and ICP algorithms, so as to construct a cumulated point cloud; Then, the cumulated and registered 3D data are filtered by exploiting the expected distribution of the acquisition error, and by estimating the resulting face manifold using a variant of the *lowess* method.Qualitative and quantitative experiments have been performed by extending the *Florence Superface* dataset with sequences of low-resolution 3D frames acquired with a Kinect camera according to an uncooperative protocol. Results of the reconstruction process of high-resolution models are evaluated by measuring the distance error between the reconstructed models and the high-resolution 3D scans used as the ground truth data of a subject's face. Results support the idea that constructing higher-resolution models from consumer depth cameras can be a viable approach to make such devices deployable in real application contexts that also include identity recognition and/or re-identification using 3D faces.

References

[1] G. Passalis, P. Perakis, T. Theoharis, and I. A. Kakadiaris, "Using facial symmetry to handle pose variations in real-world 3D face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1938–1951, Oct. 2011.

[2] H. Drira, B. Ben Amor, M. Daoudi, A. Srivastava, and R. Slama, "3D face recognition under expressions, occlusions, and pose variations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2270–2283, Sep. 2013.

[3] M. Pamplona Segundo, L. Silva, and O. Bellon, "Real-time scale-invariant face detection on range images," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Anchorage, AL, USA, Oct. 2011, pp. 914–919.

[4] F. Pala, R. Satta, G. Fumera, and F. Roli, "Multimodal person reidentification using RGB-D cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 788–799, Apr. 2016.

[5] M. Pamplona Segundo, S. Sarkar, D. Goldgof, L. Silva, and O. Bellon, "Continuous 3d face authentication using RGB-D cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, Jun. 2013, pp. 64–69.

[6] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On RGB-D face recognition using Kinect," in *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Washington, DC, USA, Sep./Oct. 2013, pp. 1–6.

[7] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna, "Using Kinect for face recognition under varying poses, expressions, illumination and disguise," in *Proc. IEEE Work. Appl. Comput. Vis. (WACV)*, Clearwater, FL, USA, Jan. 2013, pp. 186–192.

[8] R. Min, J. Choi, G. Medioni, and J.-L. Dugelay, "Real-time 3D face identification from a depth camera," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Tsukuba, Japan, Nov. 2012, pp. 1739–1742.

[9] S. Berretti, P. Pala, and A. Del Bimbo, "Face recognition by super-resolved 3D models from consumer depth cameras," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 9, pp. 1436–1449, Sep. 2014.

[10] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[11] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for ToF 3D shape scanning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 343–350.
[12] S. Liang, I. Kemelmacher-Shlizerman, and L. G. Shapiro, "3d face hallucination from a single depth frame," in *Proc. Int. Conf. 3D Vis.*, Tokyo, Japan, Dec. 2014, pp. 1–8.
[13] G. Pan, S. Han, Z. Wu, and Y. Wang, "Super-resolution of 3D face," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Graz, Austria, May 2006, pp. 389–401.
[14] S. Peng, G. Pan, and Z. Wu, "Learning-based super-resolution of 3D face model," in *Proc. IEEE Int. Conf Image Process. (ICIP)*, vol. 2. Genoa, Italy, Sep. 2005, pp. 382–385.
[15] J. Choi, A. Sharma, and G. Medioni, "Comparing strategies for 3D face recognition from a 3D sensor," in *Proc. IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Gyeongju, South Korea, Aug. 2013, pp. 19–24.
[16] M. Hernandez, J. Choi, and G. Medioni, "Laser scan quality 3-D face modeling using a low-cost depth camera," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 1995–1999.
[17] R. Newcombe *et al.*, "Kinectfusion: Real-time dense surface mapping and tracking," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Basel, Switzerland, Oct. 2011, pp. 1–10.
[18] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 343–352.
[19] S. Izadi *et al.*, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. ACM SIGGRAPH*, Vancouver, Canada, Aug. 2011, p. 1.
[20] K. Al Ismaeil, D. Aouada, T. Solignac, B. Mirbach, and B. Ottersten, "Real-time non-rigid multi-frame depth video super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 8–16.
[21] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proc. 3rd Int. Conf. 3-D Digit. Imag. Modeling*, Quebec City, QC, Canada, May 2001, pp. 145–152.
[22] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, Dec. 2010.
[23] E. Bondi, P. Pala, S. Berretti, and A. Del Bimbo, "Reconstructing high-resolution face models from Kinect depth sequences acquired in uncooperative contexts," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 07. Ljubljana, Slovenia, May 2015, pp. 1–6.
[24] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[25] J. Williams and M. Bennamoun, "Multiple view surface registration with error modeling and analysis," in *Proc. Int. Conf. Image Process.*, vol. 1. Vancouver, BC, Canada, Sep. 2000, pp. 545–548.
[26] R. Sagawa, N. Osawa, and Y. Yagi, "A probabilistic method for aligning and merging range images with anisotropic error distribution," in *Proc. 3rd Int. Symp. 3D Data Process., Vis., Transmiss.*, Chapel Hill, NC, USA, Jun. 2006, pp. 559–566.
[27] H. Hoffmann, S. Schaal, and S. Vijayakumar, "Local dimensionality reduction for non-parametric regression," *Neural Process. Lett.*, vol. 29, no. 2, pp. 109–131, Apr. 2009.
[28] P. Frasconi, L. Silvestri, P. Soda, R. Cortini, F. S. Pavone, and G. Iannello, "Large-scale automated identification of mouse brain cells in confocal light sheet microscopy images," *Bioinformatics*, vol. 30, no. 17, pp. i587–i593, Sep. 2014.
[29] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.
[30] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[31] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *J. Amer. Statist. Assoc.*, vol. 74, no. 368, pp. 829–836, 1979.
[32] R. Min, N. Kose, and J.-L. Dugelay, "KinectFaceDB: A Kinect database for face recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst.*, vol. 44, no. 11, pp. 1534–1548, Nov. 2014.
[33] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst., (BTAS)*, Washington, DC, USA, Sep./Oct. 2013, pp. 1–6.
[34] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The Florence 2D/3D hybrid face dataset," in *Proc. ACM Work. Human Gesture Behavior Understand. (J-HGBU)*, Scottsdale, AR, USA, Dec. 2011, pp. 79–80.
[35] S. Berretti, A. Del Bimbo, and P. Pala, "Superfaces: A super-resolution model for 3D faces," in *Proc. Int. Workshop Non-Rigid Shape Anal. Deformable Image Alignment (NORDIA)*, Florence, Italy, Oct. 2012, pp. 73–82.
[36] C. Xu, T. Tan, Y. Wang, and L. Quan, "Combining local features for robust nose location in 3D facial data," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1487–1494, Oct. 2006.
[37] S. Berretti, A. Del Bimbo, and P. Pala, "Sparse matching of salient facial curves for recognition of 3-D faces with missing parts," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 2, pp. 374–389, Feb. 2013.

Enrico Bondi received the master's degree in computer engineering from the University of Florence in 2014, with a thesis titled "Crowd Counting and Analysis System via Depth Sensor." He is currently a Researcher with the Media Integration and Communication Center, University of Florence. His main research interests focus on 3D face reconstruction from low-resolution scanners.

Pietro Pala received the Electronic Engineering degree and the Ph.D. degree in information and telecommunications engineering from the University of Firenze in 1994 and 1997, respectively. He is currently an Associate Professor with the University of Firenze. His research activity has focused on the use of pattern recognition models for multimedia information retrieval and biometrics. Recently, the research activity focuses on the analysis of 3D data for person recognition and human activity recognition.

Stefano Berretti is an Associate Professor with the Department of Information Engineering and Media Integration and Communication Center, University of Florence. His research interests have been mainly focused on content modeling, retrieval, and indexing of image and 3D object databases. Recent research has addressed 3D object retrieval and partitioning, 3D/4-D face and facial expression recognition, 4-D action recognition. He is an Information Director of the *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Alberto Del Bimbo is Full Professor of Computer Engineering, and Director of the Media Integration and Communication Center with the University of Florence. He was the Deputy Rector for Research and Innovation Transfer with the University of Florence from 2000 to 2006. His scientific interests are multimedia information retrieval, pattern recognition, image and video analysis, and natural human-computer interaction. He has authored or coauthored over 300 publications in some of the most distinguished scientific journals and international conferences, and is the author of the monography "Visual Information Retrieval." He is an IAPR Fellow, associate editor of several leading journals in the area of pattern recognition and multimedia, and Editor-in-Chief of the *ACM Transactions on Multimedia Computing, Communications, and Applications*.