Face Recognition by Super-Resolved 3D Models From Consumer Depth Cameras

Stefano Berretti, Pietro Pala, *Senior Member, IEEE*, and Alberto del Bimbo, *Member, IEEE*

***Abstract*— Face recognition based on the analysis of 3D scans has been an active research subject over the last few years. However, the impact of the resolution of 3D scans on the recognition process has not been addressed explicitly, yet being an element of primal importance to enable the use of the new generation of consumer depth cameras for biometric purposes. In fact, these devices perform depth/color acquisition over time at standard frame-rate, but with a low resolution compared to the 3D scanners typically used for acquiring 3D faces in recognition applications. Motivated by these considerations, in this paper, we define a super-resolution approach for 3D faces by which a sequence of low-resolution 3D face scans is processed to extract a higher resolution 3D face model. The proposed solution relies on the scaled iterative closest point procedure to align the low-resolution scans with each other, and estimates the value of the high-resolution 3D model through a 2D box-spline functions approximation. To evaluate the approach, we built—and made it publicly available—the Florence Superface dataset that collects high-resolution and low-resolution data for about 50 different persons. Qualitative and quantitative results are reported to demonstrate the accuracy of the proposed solution, also in comparison with alternative techniques.**

***Index Terms*— 3D face recognition, 3D super-resolution, 2D Box-splines, rigid registration.**

I. Introduction

In recent years, many approaches have been proposed to support person recognition by the analysis of 3D face scans. This research area is attracting an increasing interest, with several challenging issues successfully investigated, such as 3D face recognition in the presence of non-neutral facial expressions [1]–[3], occlusions [4], [5], and missing data [6], [7], to say a few. Existing solutions have been tested following well defined evaluation protocols on consolidated benchmark datasets, which provide a reasonable coverage of the many different traits and characteristics of the human face, including variations in terms of gender, age, ethnicity, hair style and occlusions due to external accessories (i.e., eyeglasses, caps, scarves, hand gestures covering part of the face, etc.) [8], [9]. The resolution at which 3D face scans are acquired changes across different datasets, but it is typically the same within one dataset. In doing so, 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 to perform recognition across scans acquired with different resolutions. This is mainly motivated by the introduction in the marketplace of a new generation of low-cost, low-resolution 4D scanning devices (i.e., 3D plus time), such as Microsoft Kinect or Asus Xtion PRO LIVE. In fact, these devices are capable of a combined color-depth (RGB-D) acquisition over time (at about 30fps), with a resolution of 18ppi at a distance of about 80*cm* from the scanning device. The spatial resolution of such devices is lower than that made possible by high-resolution 3D scanners, such as the Minolta Vivid or 3dMD. But these high-resolution scanners 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 aiming to perform face detection [10], continuous authentication [11] and recognition [12]–[14] directly from the depth data of low-resolution frames of the Kinect camera. Furthermore, 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 expected to be of lower resolution and potentially acquired with 4D cameras.

In this context, evaluating the impact on the recognition accuracy of matching low-resolution probes against high-resolution gallery scans is certainly an issue, but an even more challenging task with potentially wider applications is given by the reconstruction of one super-resolved face model out of a sequence of low-resolution depth frames acquired by a 4D scanner. In fact, constructing a higher resolution 3D model from a sequence of low-resolution 3D scans could pave the way to more versatile 3D face recognition methods deployable in contexts where acquisition of high resolution 3D scans is not convenient or even possible. For example, at a police roadblock, a police patrol has to verify the identity of a suspect individual which is stopped for a control. A 4D camera could be used to acquire a short 3D sequence from which a super-resolved model can be derived to support 3D face recognition. In a very different scenario, a user in front of a PC equipped with a 4D camera could use a super-resolution framework to construct an avatar of sufficient quality to make the individual recognizable in any virtual environment s/he uses for communication or entertainment purposes. Again, a

Manuscript received December 11, 2013; revised March 27, 2014 and June 3, 2014; accepted June 30, 2014. Date of publication July 8, 2014; date of current version July 31, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kar-Ann Toh.

The authors are with the Department of Information Engineering, University of Florence, Florence 50139, Italy (e-mail: stefano.berretti@unifi.it; pietro.pala@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.2014.2337258

4D camera could be used to acquire a short 3D sequence, and use it to produce a super-resolved face model of sufficient accuracy to support effective recognition at a check-in totem of an airport, rather than in front of an ATM or at a bank entrance.

Based on these premises, in this work we aim to provide an effective *super-resolution* approach specifically tailored for 3D faces that can help reducing the gap between low- and high-resolution acquisitions in several applicative scenarios.

A. Related Work

Formerly introduced for 2D still images, super-resolution aims at recovering one high-resolution image from a set of low-resolution images possibly altered by noise, blurring or geometric warping [15]–[19]. Super-resolution approaches proposed in the specific context of 3D data can be grouped in two classes: Approaches that apply the super-resolution in the 2D space and then use multiple super-resolved 2D images to reconstruct a super-resolved 3D model [20]; And approaches that operate directly in the 3D space by applying a super-resolution approach to 3D data [21]–[24]. In the following, we focus our review on the methods of the second class.

The approach proposed in [23] is conceived to operate on data provided by time-of-flight cameras. These data are upsampled and denoised by using information from a high-resolution image of the same scene that is taken from a viewpoint close to the depth sensor. The denoising module exploits the relations between depth and intensity data, such as the joint occurrence of depth and intensity edges, and smoothness of geometry in areas of largely uniform color. The approach proposed in [22] also targets processing of data provided by time-of-flight cameras. However, this solution relies on an energy minimization framework that explicitly takes into account the characteristic 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. In general, approaches that deal with 3D data representing multiple objects in complex scenes focus on the relevance of accurate reconstruction in correspondence to discontinuities of the depth value that are associated with object boundaries. However, this issue is less relevant if the 3D data represent a single object with smooth surface such as a face. Previous work that focus on super-resolution of 3D faces are reported in [21] and [24]. In [21], a learning module is trained on high resolution 3D face models so as to learn the mapping between low-res data and high-res data. Given a new low-res face model the learned mapping is used to compute the high-res face model. Differently, in [24] the super-resolution process is modeled as a progressive resolution chain whose features are computed as the solution to a MAP problem. In both cases, the adopted framework is validated on synthetic data, that is, using high-resolution 3D face models, down-sampling and adding random noise to these models and then adopting the super-resolution framework to reconstruct a super-resolved 3D face that is compared to the original high-resolution 3D face model. However, the fact that low resolution models are artificially derived from high resolution ones may bias the effectiveness of learning and MAP modeling. Differently, an unbiased estimation of the accuracy of the super-resolution should use truly real data: Comparing a 3D face model acquired with a high-resolution scanner to a super-resolved 3D face model reconstructed from facial data acquired by a low resolution scanner.

Methods in [25], [26], and [27] approach the problem of noise reduction in depth data by fusing the observations of multiple scans to construct one denoised scan. In [25], 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 object. Later, dynamic interaction has been added to the system in [28], where camera tracking is performed on a static background scene and the foreground object is tracked independently of camera tracking. Aligning all depth points to the complete scene model from a large environment (e.g., a room) provides very accurate tracking of the camera pose and mapping [25]. However, this approach is targeted to generic objects in internal environments, rather than to faces. In [26], a 3D face model with an improved quality is obtained by a user moving in front of a low-cost, low resolution depth camera. The face is represented in cylindrical coordinates, which enables efficient filtering operations. 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 (Graphics Processing Unit) implementation of the ICP (Iterative Closest Point) algorithm. Temporal and spatial smoothing of the incrementally refined model are also performed. The approach is validated by comparing quantitatively the obtained 3D model to one produced by high-resolution laser scanning. This approach is used in [27] 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: In the first scenario, the probe is given by a single raw depth frame (1F), which is compared against a gallery comprising, respectively, a frame (1F), multiple frames (NF) and a reconstructed 3D face model (3D) for each subject; In the second scenario, a reconstructed 3D face model (3D) is used as probe, which is compared against a gallery of reconstructed 3D face models (3D). The reported analysis on a small dataset of 10 subjects shows that the 3D-3D and the 1F-NF scenarios provide better results compared to the baseline 1F-1F approach. Although the method is not conceived to increase the resolution of the reconstructed model with respect to the individual frames, it supports the idea that aggregating multiple observations enhances the signal to noise ratio, thus increasing the recognition results with respect to the solution where a single frame is used.

B. Our Method and Contribution

In this paper, we present an original solution to derive one super-resolution 3D face model from the low-resolution depth frames of a sequence acquired through a depth scanner (a Microsoft Kinect camera is used). The proposed approach relies on scattered data approximation techniques and operates in three main processing steps: *(i)* First, for each depth frame of the sequence, the region containing the face is automatically detected and cropped; *(ii)* Then, the cropped face of the first frame of the sequence is used as reference and all the faces cropped from the other frames are aligned to the reference; *(iii)* Finally, data obtained by aggregating these multiple aligned observations are resampled at a higher resolution and approximated using 2D-Box splines.

To validate the proposed approach and estimate the accuracy of the reconstructed super-resolved models, the *The Florence Superface* v2.0 dataset has been constructed. For each individual, the dataset includes one sequence of depth frames acquired through a Kinect scanner as well as one high-resolution face scan acquired through a 3dMD scanner. In this way, the accuracy of the reconstructed super-resolved model can be quantitatively measured by comparing the reconstructed model to the corresponding high-resolution scan.

In summary, the main contributions of this paper are:

- A complete approach to reconstruct a super-resolved 3D face model from a sequence of low-resolution depth frames of the face;
- A thorough evaluation of the proposed super-resolution approach to demonstrate that: *i)* It produces a super-resolved 3D model rather than just a denoised one; *ii)* Use of the reconstructed super-resolved 3D face model in recognition experiments improves face recognition accuracy compared to the case in which a low resolution depth frame is used;
- A public heterogeneous 3D face dataset, which includes low-resolution depth sequences of the face as well as high-resolution 3D face scans of the same subjects.

Preliminary ideas and results related to the proposed method were first reported in [29]. With respect to our previous work, we propose a new formulation of the super-resolution approach, which now relies on the family of 2D Box-splines approximating functions. In addition, we demonstrate that our solution results in a super-resolved model, rather than producing just a denoised surface. Furthermore, we completely revised the experimental evaluation of the proposed approach that now is performed on the *The Florence Superface* v2.0 dataset and includes new experiments, where the super-resolved models are used as probes to perform 3D face recognition against high-resolution gallery scans.

The paper is organized as follows: The problem statement and the basic notation are defined in Sect. II, together with the face detection, cropping and alignment operations performed on individual frames of 3D depth sequences; The super-resolution approach based on facial data approximation is described and validated in Sect. III. Experimental results are reported and discussed in Sect. IV, where first the error between super-resolved models and high-resolution scans is computed, then performance measures of the recognition accuracy using super-resolved models as probes against high-resolution gallery are reported. Finally, discussion and conclusions are given in Sect. V.

II. Problem Statement and Processing of Depth Sequences

In the literature, the super-resolution process is typically formalized on 2D still images as an inverse problem: The low resolution images are the observations from slightly different viewpoints of a high resolution image, the underlying scene. It should be noticed that the relative motion between the scene and the camera is a necessary prerequisite to guarantee that pixels in the low-resolution images represent new samples of the patches in the observed scene. No improvement on resolution—if any, only in terms of denoising—would be possible from images deriving from a fixed camera observing a static scene.

The proposed approach targets the reconstruction of a *depth image* of the face (*image* for short), which shows both super-resolution and denoising, starting from a sequence of low-resolution *depth frames* (*frames* in the following). To simplify the notation and without loss of generality we assume that each frame is defined on a regular low-resolution grid $\Omega = [1, \ldots, N] \times [1, \ldots, N]$. The high-resolution image is defined on a regular high-resolution grid $\Sigma = [1, \ldots, M] \times [1, \ldots, M]$, being $\zeta = M/N$ the *resolution gain*. The forward degradation model, describing the formation of low-resolution frames from a high-resolution image, can be formalized as follows:

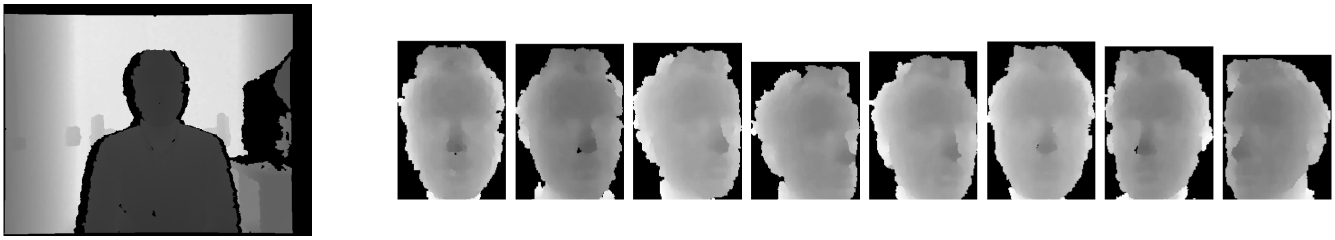
$$X_L^{(k)} = P_k(X_H) \quad k = 1, \ldots, K, \tag{1}$$

being $\left\{X_L^{(k)}\right\}_{k=1}^{K}$ the set of K low-resolution frames, X_H the high-resolution image, and P_k the operator that maps the high-resolution image onto the coordinate system and sampling grid of the k-th low-resolution frame. The mapping operated by P_k accounts for four main factors: *(i)* The geometric transformation of X_H to the coordinates of the k-th low-resolution frame $X_L^{(k)}$; *(ii)* The blurring effect induced by the atmosphere and camera lens; *(iii)* Down-sampling; and *(iv)* Additive noise.

The coordinate system of the high-resolution image X_H is aligned to the coordinate system of the first low-resolution frame $X_L^{(1)}$ of the sequence, which is used as *reference*. That is, the computation of the geometric transformation that maps the coordinate systems of subsequent low-resolution frames is operated by registering the low-resolution frames to the first frame of the sequence. This is obtained through an iterative procedure, which is applied using the cropped region of the face detected in each frame (details on face detection, cropping and frame registration are given in the remaining part of this Section). The cumulated data obtained by the alignment of data from the sequence of low-resolution frames to data in the first low-resolution frame $X_L^{(1)}$, represent a point cloud in the 3D space. Points of the cloud are regarded as observations of the value of the high-resolution image X_H. The estimate of this high-resolution image is formalized as the solution of a scattered data approximation problem, as described in Sect. III.

A. Depth Frames Acquisition and Face Cropping

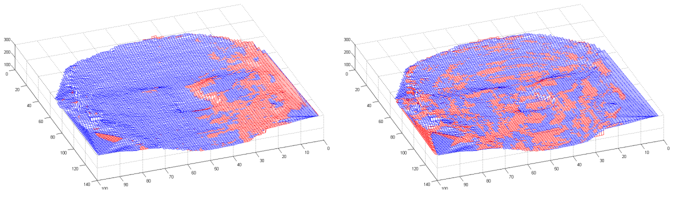
In our approach, low-resolution frames are acquired by a Kinect scanner placed in front of a subject sitting at a distance



(a) (b)

Fig. 1. (a) Sample depth frame acquired by the scanning device. (b) Some cropped faces from the sequence of acquired frames. It can be observed as the pose of the face varies from frontal to left and right side.

of about 80*cm* from the scanning device. It is assumed that the sequence of acquired frames represents the subject while s/he is slightly rotating the head to the left and right around the vertical axis (the neck). In Fig. 1(a) a sample frame is shown out of a sequence of depth frames acquired by the scanner. Acquired frames are processed in order to crop each frame in correspondence to the face of the subject. For this purpose, the Face Tracking function supported by the device SDK has been used. Some representative frames output by the face cropping module for a sample sequence are shown in Fig. 1(b).



(a) (b)

Fig. 2. Facial data acquired in two sample frames (one shown with red and one with blue colors) before (a) and after (b) the application of the ICP procedure. (Figure best viewed on soft-copy version.)

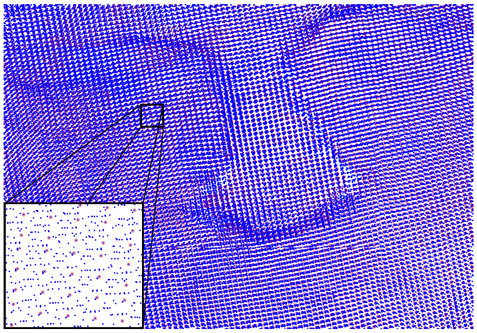


Fig. 3. Result of the alignment of points of the frames of a sample sequence (blue points) to the points of the reference frame (red circles). (Figure best viewed on soft-copy version.)

B. Registration of Subsequent Frames

Computation of the geometric transformation that aligns low-resolution frames to a common reference system is performed through a variant of the Iterative Closest Point (ICP) procedure [30], which jointly estimates the 3D rotation and translation parameters as well as the scaling one [31]. Let $\mathbf{x}_i^{(k)}$ be the 3D coordinates (x, y and the depth value z) of the i-th facial point in the k-th frame $X_L^{(k)}$. Registration of facial data represented in $X_L^{(k)}$ to data represented in the reference frame $X_L^{(1)}$ is obtained by computing the similarity transform (translation, rotation and scaling) that best aligns the transformed data to the data in the reference frame, that is:

$$\min_{\mathbf{R},\mathbf{S},\mathbf{t},p}\left(\sum_{i=1}^{\left|X_L^{(k)}\right|}\left\|\mathbf{R}\cdot\mathbf{S}\cdot\mathbf{x}_i^{(k)}+\mathbf{t}-\mathbf{x}_{p(i)}^{(1)}\right\|\right), \qquad (2)$$

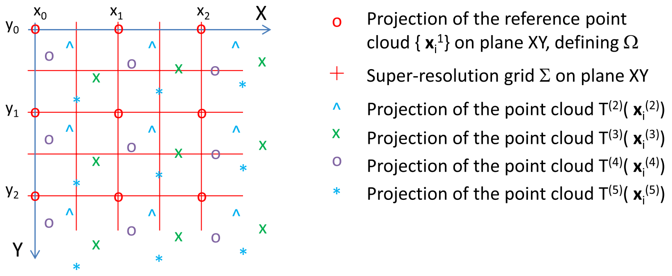
being $\mathbf{R}$ an orthogonal matrix, $\mathbf{S}$ a diagonal scale matrix, $\mathbf{t}$ a translation vector, $|.|$ the cardinality of a set, and $p : \left\{1,\ldots,\left|X_L^{(k)}\right|\right\} \mapsto \left\{1,\ldots,\left|X_L^{(1)}\right|\right\}$ a function that maps indexes of facial points across the k-th and the 1-st frames. The solution of Eq. (2), namely $\mathbf{R}^k, \mathbf{S}^k, \mathbf{t}^k$, is computed according to the procedure described in [31]. To simplify the notation, the overall effect of $\mathbf{R}^k, \mathbf{S}^k, \mathbf{t}^k$ on the points $\mathbf{x}_i^{(k)}$ acquired at the k-th frame is indicated as $\mathbf{T}^k(\mathbf{x}_i^{(k)})$. Figure 2 shows facial data acquired in two sample frames before and after the application of the ICP procedure.

The ICP algorithm usually requires an appropriate initialization to yield high registration accuracy. For this purpose, alignment of the data in a generic frame $X_L^{(k)}$ to the data in the reference frame $X_L^{(1)}$ is obtained by first applying to $X_L^{(k)}$ the transformation computed for the previous frame $X_L^{(k-1)}$. In this way, the transformation of the (k-1)-th frame is used to predict the transformation of the k-th frame, and ICP is used for fine registration of the prediction against the reference.

The result of aligning a frame sequence to the first frame (reference) of the sequence is summarized in Fig. 3. In this plot, the points of the reference frame are shown with red circles, whereas the points of the other frames of the sequence after alignment are reported with blue points. These scattered and irregularly distributed points are the input to the super-resolution procedure.

III. Super-Resolution Approach

Based on the procedure described so far, points of the frames $X_L^{(k)}$, $k = 2,\ldots,K$ are aligned to the data in the first frame $X_L^{(1)}$, used as reference. The first frame implicitly defines a 3D Cartesian reference system (X, Y, Z) and its associated sampling grid $\mathbf{\Omega} = \{1,\ldots,N\}\times\{1,\ldots,N\}$. In this reference system, the acquired values of the facial surface are regarded as samples of a continuous function $f^1(x, y)$



x_0
x_1
x_2
X
y_0
y_1
y_2
Y
o Projection of the reference point cloud { $\mathbf{x}_i^1$} on plane XY, defining Ω
+ Super-resolution grid Σ on plane XY
^ Projection of the point cloud $T^{(2)}(\mathbf{x}_i^{(2)})$
x Projection of the point cloud $T^{(3)}(\mathbf{x}_i^{(3)})$
o Projection of the point cloud $T^{(4)}(\mathbf{x}_i^{(4)})$
* Projection of the point cloud $T^{(5)}(\mathbf{x}_i^{(5)})$

Fig. 4. The original sampling grid $\mathbf{\Omega}$ is defined by the x- and y-coordinates of points acquired in the reference frame. Data acquired in subsequent frames are subject to ICP alignment and their projections on the XY plane do not fit the sampling grid $\mathbf{\Omega}$. The super-resolution grid Σ is obtained by oversampling the original grid $\mathbf{\Omega}$ by a factor ζ ($\zeta = 2$ in the Figure). (Figure best viewed on soft-copy version.)

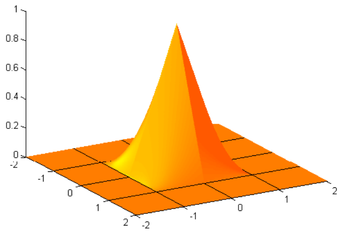


Fig. 5. Plot of the 2D base function $B_{0,0}(x, y)$. Points of the lattice correspond to 2D points with integer coordinates. The base function has non-zero values only inside the region $[-1, 1] \times [-1, 1]$ that results from the union of the four cells with one vertex on the point $x_0, y_0 = (0, 0)$.

at the points of the grid $\mathbf{\Omega}$. As described in Sect. II-B, these sampled values can be equivalently considered a point cloud $\{\mathbf{x}_i^{(1)}\} = \{(x_i, y_i, f^1(x_i, y_i))\}$ in the 3D Cartesian reference system (X, Y, Z). Data acquired in the second scan $\{\mathbf{x}_i^{(2)}\} = \{(x_i, y_i, f^2(x_i, y_i))\}$ are expressed on the sampling grid $\mathbf{\Omega}$, but they represent the face under the effect of a small rotation angle (the user rotates the head during the data acquisition process). ICP is used to determine the best similarity transform $\mathbf{T}^{(2)}$ that aligns the point cloud $\{\mathbf{x}_i^{(2)}\}$ to the point cloud of the reference frame $\{\mathbf{x}_i^{(1)}\}$. Under the effect of this similarity transform, points of the point cloud $\{\mathbf{x}_i^{(2)}\}$ are transformed to $\mathbf{T}^{(2)}(\mathbf{x}_i^{(2)})$. Therefore, the $x-$ and $y-$coordinates of these points are no longer aligned to the sampling grid $\mathbf{\Omega}$. In general, this applies to all the point clouds acquired after the reference one.

A graphic representation of this general scenario is provided in Fig. 4. This evidences that the $x-$ and $y-$coordinates of points acquired in the reference frame implicitly define the original sampling grid $\mathbf{\Omega}$, and data acquired in subsequent frames, after ICP alignment, are scattered on the original grid. The super-resolution grid $\Sigma = \{1, \ldots, M\} \times \{1, \ldots, M\}$ is defined by oversampling the original grid $\mathbf{\Omega}$ by a factor $\zeta = M/N$, representing the resolution gain. If only the data of the reference frame were available, estimation of the value of the face surface in correspondence to points of the super-resolution grid Σ would rely on the interpolation of available points $\{(x_i, y_i, f^1(x_i, y_i))\}$, yielding no true super-resolution. Differently, the availability of data from subsequent frames enables the estimation of the value of the face surface using a higher number of observations compared to the data of the reference frame. This is obtained by gathering all data from available frames and approximating these data through a function $\Gamma(x, y)$ defined on the super-resolution grid Σ.

Let us consider the set of points obtained by cumulating the aligned observations across different frames and denote it as:

$$\mathcal{O} \equiv \{Px_l, Py_l, Pz_l\}_{l=1}^{L} = \bigcup_{k=1}^{K} \left\{\mathbf{T}^{(k)}(\mathbf{x}_i^{(k)})\right\}, \tag{3}$$

being $L = \sum_{k=1}^{K} \left|X_L^{(k)}\right|$. To estimate the values of the face surface, observations in $\mathcal{O}$ are approximated by a function $\Gamma(x, y)$ that is expressed through the 2D Box-splines model [32]–[34]. Following this model, the function $\Gamma(x, y)$ is a weighted sum of Box-splines obtained by shifting a 2D base function $B_{0,0}(x, y)$ with local support. Given a 1D regular lattice $\{x_{-M}, \ldots, x_{-1}, x_0, x_1, \ldots, x_M\}$, with $\Delta = x_{i+1} - x_i$, the 1D first degree (C^0 continuity) base function $b_0(t)$ is defined as:

$$b_0(t) = \begin{cases} 0 & if \quad t \in (-\infty, x_{-1}] \\ \frac{t - x_{-1}}{\Delta} & if \quad t \in (x_{-1}, x_0] \\ \frac{x_1 - t}{\Delta} & if \quad t \in (x_0, x_1] \\ 0 & if \quad t \in (x_1, \infty) . \end{cases} \tag{4}$$

The shifted copy of the base function, centered on the generic node x_i of the lattice is computed as $b_i(t) = b_0(t - x_i)$. Extension of this framework to the 2D case is operated by considering a 2D lattice $\{x_i, y_j\}$ $i, j = 0, \ldots, M$, whose elements correspond to the nodes of the grid Σ. In this case, the 2D base function $B_{0,0}(x, y)$ is computed as the tensor product of the 1D base function:

$$B_{0,0}(x, y) = b_0(x) b_0(y). \tag{5}$$

The shifted copy of the base function, centered on the generic node x_i, y_j of the lattice is computed as $B_{i,j}(x, y) = b_i(x) b_j(y)$. Functions $B_{i,j}(x, y)$ are continuous, with local support, and are zero for all points (x, y) such that $\|(x, y) - (x_i, y_j)\|_\infty > \Delta$. As an example, the plot of the 2D base function $B_{0,0}(x, y)$ is shown in Fig. 5.

The approximating function $\Gamma(x, y)$ is expressed as a weighted combination of base functions centered at nodes of the super-resolution grid Σ:

$$\Gamma(x, y) = \sum_{i,j} w_{i,j} B_{i,j}(x, y). \tag{6}$$

Values of the weights $w_{i,j}$ are determined so as to yield the best approximation of $\Gamma(x, y)$ to the point cloud. In order to determine the values of these weights, two types of constraints are considered targeting the fit of $\Gamma(x, y)$ to the points and the regularity of $\Gamma(x, y)$, in terms of continuity and derivability. In the ideal case, $\Gamma(x, y)$ would fit all the points. This constraint is expressed by L equations of the form:

$$\Gamma(Px_l, Py_l) = Pz_l \quad l = 1, \ldots, L, \tag{7}$$

being L the overall number of points obtained by registering all the K frames of the sequence (see Eq. (3)). Due to the form of the basis functions (Eqs. (4)-(5)), $\Gamma(x, y)$ is continuous

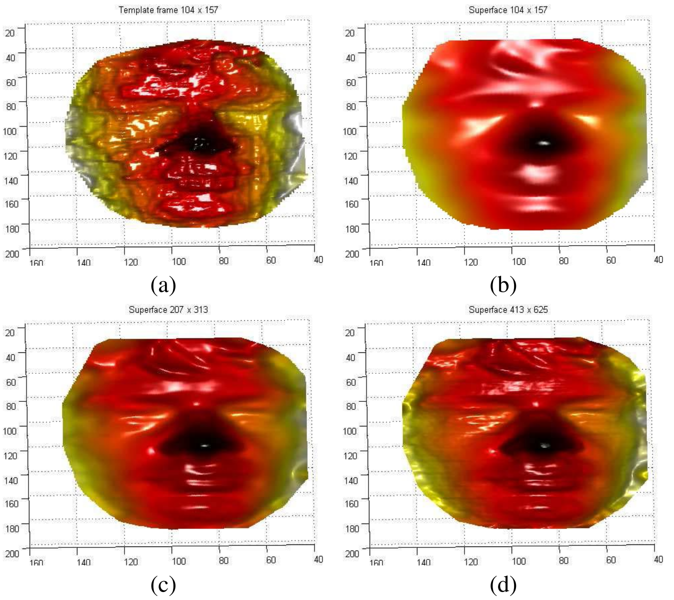


Fig. 6. (a) Reference frame of a sequence. Three models reconstructed at different resolutions are reported: (b) 104×157, same resolution as the original (just denoising); (c) 207 × 313; (d) 413 × 625.

everywhere. Since $\Gamma(x, y)$ is not derivable in correspondence to the points of the lattice $\{x_i, y_j\}$, its smoothness is forced by the following additional set of equations:

$$\left.\frac{\partial^{+}\Gamma(x,y)}{\partial x}\right|_{x_i,y_j} = \left.\frac{\partial^{-}\Gamma(x,y)}{\partial x}\right|_{x_i,y_j}$$
$$\left.\frac{\partial^{+}\Gamma(x,y)}{\partial y}\right|_{x_i,y_j} = \left.\frac{\partial^{-}\Gamma(x,y)}{\partial y}\right|_{x_i,y_j} \quad i,j = 1,\ldots,M-1. \quad (8)$$

The left and right partial derivatives of Eq. (8) can be obtained analytically. In fact, it can be easily shown that combination of Eqs. (4)-(5) and (8) yields to the following expressions:

$$-w_{i,j} + w_{i+1,j} = -w_{i-1,j} + w_{i,j}$$
$$-w_{i,j} + w_{i,j+1} = -w_{i,j-1} + w_{i,j} \qquad i,j = 1,\ldots,M-1. \quad (9)$$

These define a set of $(M-1)^2$ equations that combined with the L equations of Eq. (7) represent a system of linear equations in the M^2 variables $w_{i,j}$. Values of the variables $w_{i,j}$ are computed by resolving a least-squares fit, which minimizes the sum of the squares of the deviations of the data from the model.

A. Super-Resolution Gain

In the following, we show that the proposed solution results in a super-resolved surface, rather than just a surface denoising. We start by showing in Fig. 6(b)-(d) the reconstruction of a sample face at different resolutions, respectively, 104 × 157, 207 × 313 and 413 × 625. In the same Figure, the plot in (a) shows the reference frame of the sequence.

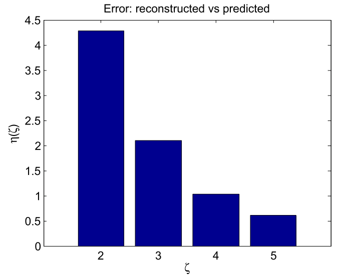


Fig. 7. Values of $\eta(\zeta)$ measure the error between the model reconstructed through the proposed super-resolution approach at the resolution gain ζ, and the prediction (by bilinear interpolation) based on the model reconstructed at the resolution gain ζ-1.

Although, in theory, the resolution gain can be set arbitrarily, the interest lies in the identification of the highest value of the *real* resolution gain, beyond which the amount of information encoded in the reconstructed surface does not change: two reconstructions of a surface at two different resolutions encode the same information if the reconstruction at the higher resolution can be obtained by resampling and interpolation of the reconstruction at the lower resolution. For this purpose, we compare results of the proposed super-resolution approach with those obtained through resampling and interpolation of data at the original resolution. Assuming $\Omega = [1,\ldots,N] \times [1,\ldots,N]$ be the original sampling grid and $\Sigma = [1,\ldots,M] \times [1,\ldots,M]$ the super-resolved one, we measure the difference between the super-resolved model reconstructed on the grid Σ and the predicted model obtained by reconstructing the face model on the original grid Ω and then increasing the resolution by resampling up to Σ and predicting values at the new grid points by bilinear interpolation. More formally, let F_ζ be the super-resolved model at a resolution $M = \zeta N$. Let $\mathcal{R}(\cdot)$ be the operator that resamples an image by bilinear interpolation, doubling the size of the input grid on both the x and y axis. The ratio η measures the mean error between the predicted and the super-resolved model:

$$\eta(\zeta) = \frac{\sum_{i,j} \left|\mathcal{R}(F_{\zeta-1}) - F_\zeta\right|}{\zeta^2 N^2}. \quad (10)$$

At the lowest value of the resolution gain, $\zeta = 2$, $F_{\zeta-1}$ is the reconstruction of the facial surface at the original resolution. Resampling this surface by bilinear interpolation yields $\mathcal{R}(F_{\zeta-1})$ whose resolution is twice the original. F_ζ is the output of the super-resolved facial surface at a resolution twice the original one. Values of $\eta(\zeta)$ are expected to decrease for increasing values of ζ. This is confirmed by the plot of Fig. 7, showing the values $\eta(\zeta)$ for $\zeta \in \{2,\ldots,5\}$. For $\zeta = 2$ the error is computed between the bilinearly interpolated reference frame and the super-resolved model at a resolution twice the original one; For increasing values of ζ, the difference between the predicted and the reconstructed models decreases showing that the higher the resolution, the lower is the information truly added by the super-resolved model compared to the information predicted by interpolation.

IV. Experimental Results

In the following, we report results of the experiments carried out to evaluate the proposed super-resolution approach. In the reported analysis, the following aspects have been addressed:

- *Accuracy of the super-resolution reconstruction*. The reconstruction error is computed between the super-resolved models and the corresponding high-resolution scans of the same subjects. Results reported in Sect. IV-B show that the super-resolved models are closer than the low-resolution scans to the corresponding high-resolution scans;
- *Reconstruction under different conditions*. The accuracy of the reconstructed model can be affected by several factors, the most notables being the choice of the reference frame and the change of facial expression during acquisition. In Sect. IV-C, we investigated the effects of these two factors on the accuracy of the super-resolution approach;
- *Comparative evaluation*. The accuracy of the super-resolved models obtained with the proposed approach has been compared, both qualitatively and quantitatively, against two alternative solutions (Sect. IV-D);
- *Accuracy of face recognition*. The reconstructed facial models have been used to perform person identification with respect to a gallery of high-resolution scans. Results in Sect. IV-E show that the use of super-resolved models paves the way to face recognition using consumer depth cameras. This is motivated by a clear improvement of the recognition accuracy when compared to the recognition performed using the low-resolution frames.

The study reported hereafter has been carried out on the *The Florence Superface* dataset [29][1]. The main features of the dataset are summarized below.

A. The Florence Superface Dataset

Very few datasets exist for face analysis from consumer cameras like Kinect (see for example the recently released EURECOM Kinect Face Dataset [35], or the The 3D Mask Attack Database (3DMAD) specifically targeted to detect face spoofing attacks [36]). However, to the best of our knowledge, no public dataset exists that provides, at the same time, sequences of low resolution face scans acquired with 3D consumer cameras and high resolution 3D scans of the same subjects[2]. So, to overcome the lack of appropriate benchmark collections and test our super-resolution approach, we constructed a proprietary dataset, which is released for free to the research community for comparative evaluations.

The *The Florence Superface* dataset (UF-S) has been collected from summer 2012 to spring 2013 at Media Integration and Communication Center of University of Florence, with the aim to support 3D face analysis across scans acquired with different devices at different resolutions. The version 1.0 of the dataset included 20 subjects and was released in October 2012 [29]. As part and further contribution of this work, we release the version 2.0 of the UF-S that now includes 50 subjects. In particular, for each person enrolled in the dataset, we captured in the same session:

- A 3D high-resolution face scan acquired with the *3dMD* scanner [37]. The scan comprises a 3D face mesh with about 40,000 vertices and 80,000 facets, and a texture stereo image with a resolution of 3341 × 2027 pixels. The geometry of the mesh is highly accurate with an average RMS error of about 0.2mm or lower, depending on the particular pre-calibration and configuration. All the high-resolution 3D scans are provided in OFF and VRML format (with the texture image);
- A video sequence acquired with the *Kinect* camera. Videos are captured so that the person sits in front of the camera with the face at an approximate distance of 80cm from the sensor. During acquisition, the subject is asked to rotate the head around the yaw axis so that both the left and right side of the face are exposed to the camera sensor. This results in video sequences lasting approximately 10 to 15 seconds on average, at 30fps. Each video is released as a sequence of frames in PNG format with 16 bits gray scale (for depth) and 24 bits color (for RGB). In both the cases, the size of the image frames is 640 × 480.

The 3D high-resolution scans and the Kinect video sequences are made available in the form produced by the sensors, without any processing or annotation. Figure 8 shows samples of the raw data acquired for two subjects. We point out that in the super-resolution approach described in this paper the RGB frames are not used. However, the dataset also includes the RGB data as in some future work these could provide additional clues for inter-frame data registration.

B. Accuracy of the Super-Resolution Models

This experiment evaluates the accuracy of the reconstructed 3D super-resolution model with respect to the 3D high-resolution scan of a same subject. In addition, in order to highlight the improved quality of the super-resolution model with respect to the original depth frames captured by the Kinect, we also considered the error between the first depth frame of a sequence (*reference frame*) and the 3D high-resolution scan. Choosing the first frame of a sequence as *reference frame* is motivated by the fact that at the beginning of the acquired video sequences, persons sit in front of the camera looking at it, so that just a few areas of the face are not visible to the sensor due to self-occlusion effects.

All the subjects in the UF-S v2.0 dataset have been used in the experiments. In particular, for each subject we considered: The high-resolution scan; The super-resolution (reconstructed) model; and the low-resolution scan (this latter obtained from the reference frame of the depth sequence). In all these cases, the 3D facial data are represented as a mesh and cropped using a sphere of radius 95*mm* centered at the nose tip (the approach in [38] is used to identify the nose tip). To measure the error between the high-resolution scan and the super-resolution model of the same subject, they are first aligned through the ICP registration algorithm [39]. Then, for each point of the super-resolution model its distance to the closest point in

[1] The *Florence Superface* dataset, available at: http://www.micc.unifi.it/datasets/4d-faces/

[2] The term “scan” is used in the following to refer to acquisitions performed with the high-resolution scanner or with the Kinect sensor (in this latter case, we also use the term “frame” as well). The term “model” is reserved instead to the 3D model obtained from the proposed super-resolution approach.

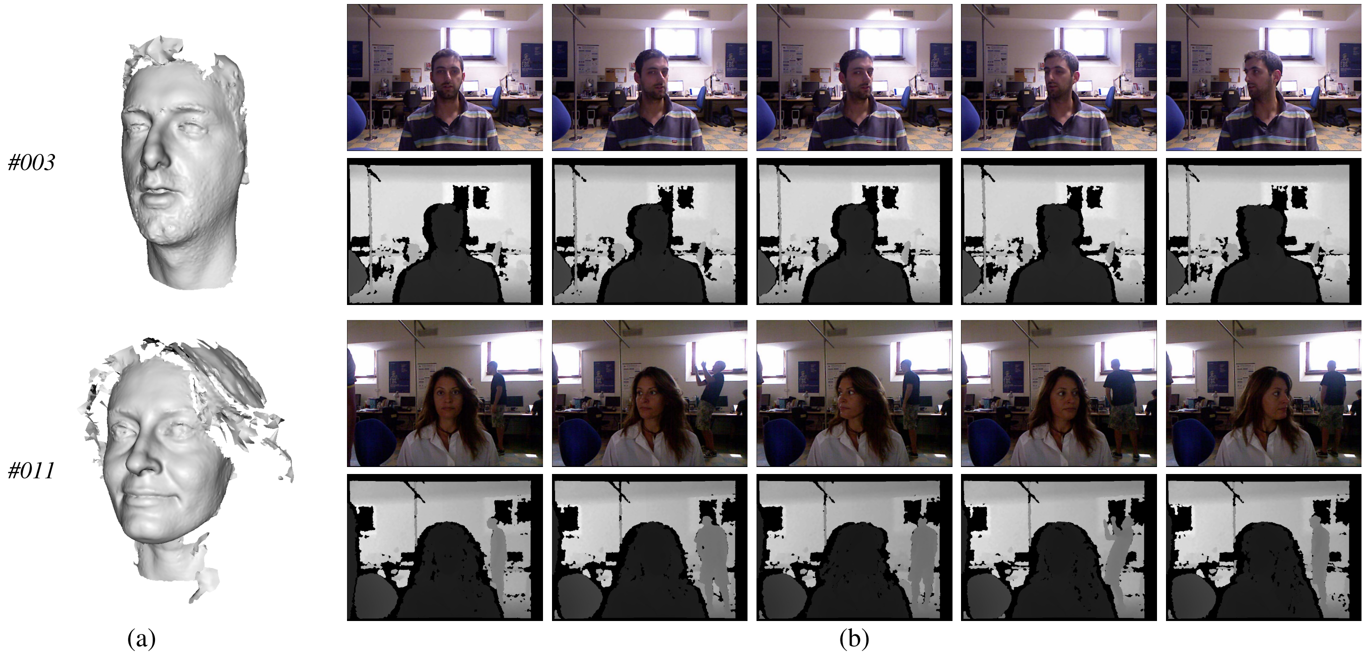


Fig. 8. *Florence Superface* dataset: (a) Sample scans acquired with the 3D high-resolution scanner (subject #003 and #011); (b) RGB-D frames sampled from the Kinect video sequences of the subjects shown in (a). It can be observed that the head pose changes from frontal to left and right side, so that a large part of the face is exposed to the sensor.

the high-resolution model is computed to build an error-map. As an example, Fig. 9 shows for some representative subjects (subject #009, #010, #011, #014, #016 and #019), one column per subject, the cropped 3D mesh of the reference frame, the super-resolution model, the high-resolution scan and the error-map between the super-resolution model and the high-resolution scan (after alignment).

a) Distance measure: The error maps, such as those reported in Fig. 9, do not directly provide a distance value to represent the dissimilarity between two models: Inter-vertex distance measures have to be combined into a distance function operating on two meshes that can have different resolutions. To represent the average error of the reconstructed models and reference frames with respect to high-resolution scans, the *Root Mean Square Error* (RMSE) between two surfaces S and S' is computed considering the vertex correspondences defined by the ICP registration, which associates each vertex $p \in S$ to the closest vertex $p' \in S'$:

$$RMSE(S, S') = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - p_i')^2}, \tag{11}$$

being N the number of correspondent points in S and S'.

Results obtained using this distance measure are summarized in Table I. In particular, we reported the average values for the *RMSE* computed between the high-resolution scan and, respectively, the super-resolution model and the reference scan. On the one hand, values in Table I measure the magnitude of the error between the super-resolution model and the high-resolution scan of same subjects; On the other, they give a quantitative evidence of the increased quality of the super-resolution model with respect to the reference scan. This latter result is indeed an expected achievement of the proposed approach, since the super-resolution models combine information of several frames of a sequence. However, it is interesting to note the substantial decrease of the error with respect to the reference frame, as can be noted by looking at the relative variation of the distance measure when passing from the reference scan to the super-resolution model (last row of Table I).

TABLE I

Average Distance Measure Computed Between the 3D High-Resolution Scan and, Respectively, the Super-Resolution Model and the Reference Scan of Each Subject. The Relative Variation of the Distance Values Is Also Reported

	average distance $RMSE$
reference vs. high-res	1.48
reconstructed vs. high-res	1.16
% variation	-21.6%

Measures in Table I provide an indication of the relative improvement of the reconstructed model vs. the reference frame. To better understand the actual improvement, it is worth considering the value of the average inter-subject distance between any two high-resolution scans of different subjects. Results of this analysis are reported in Table II that also shows the relative variation of the intra-subject distance values of Table I compared to the inter-subject high-resolution distance values. It can be noticed that compared to the average inter-subject distance, the accuracy of the super-resolution models is considerably higher than the accuracy of the reference scans.

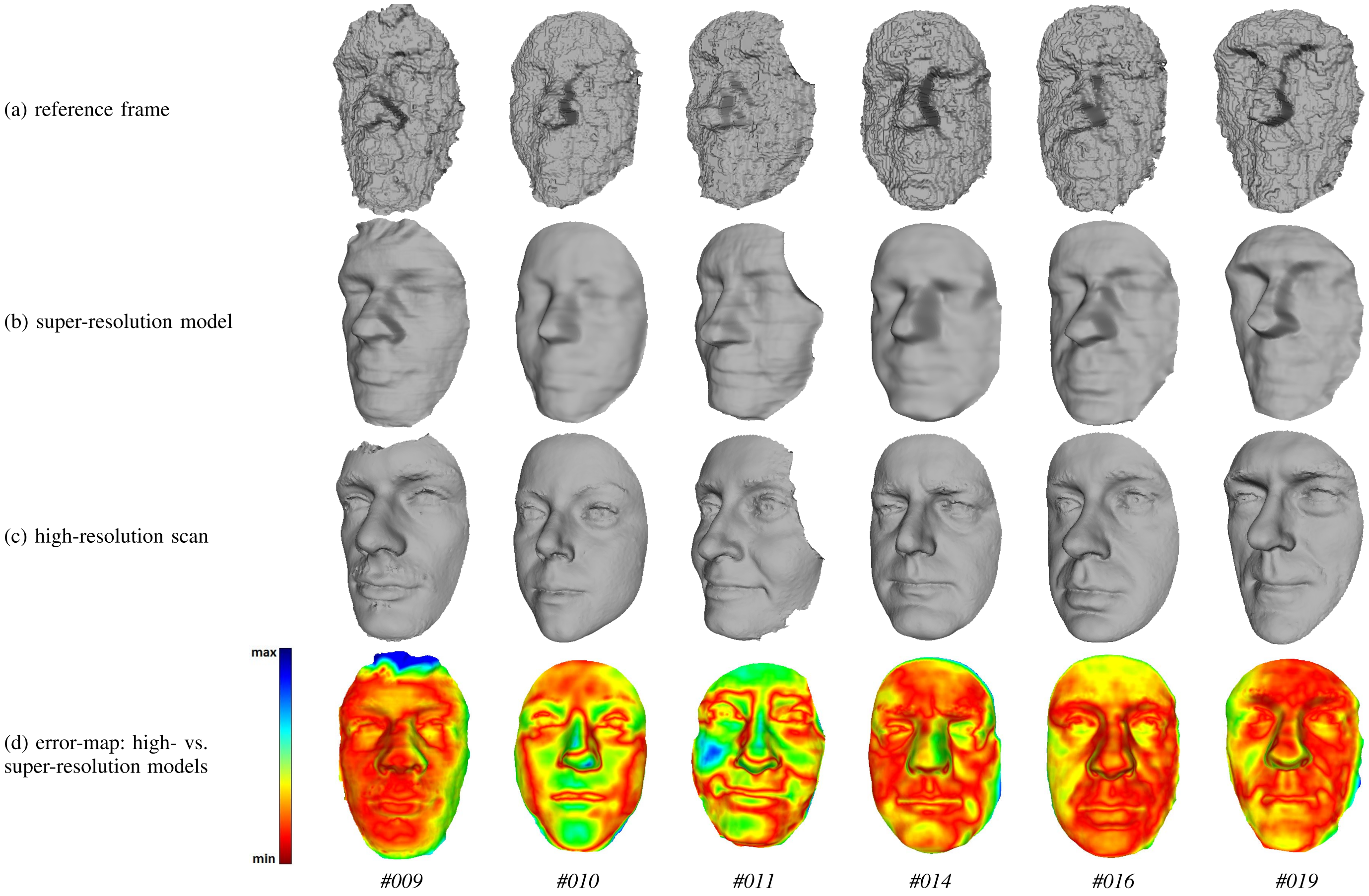

(a) reference frame
(b) super-resolution model
(c) high-resolution scan
(d) error-map: high- vs. super-resolution models
max
min
#009
#010
#011
#014
#016
#019

Fig. 9. Acquired/processed scans of the UF-S dataset. Each column corresponds to a different sample subject and reports: (a) The low resolution 3D scan of the reference frame; (b) the super-resolution 3D model; (c) the high-resolution 3D scan. The error-map in (d) shows, for each point of the super-resolution 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).

TABLE II
AVERAGE DISTANCE MEASURE COMPUTED BETWEEN ANY TWO HIGH-RESOLUTION SCANS OF DIFFERENT SUBJECTS. THE RELATIVE VARIATION OF THE INTRA-SUBJECTS DISTANCE VALUES LISTED IN TABLE I IS ALSO REPORTED

	average distance $RMSE$
high-res vs. high-res	1.42
% variat. *reference*	+4.2%
% variat. *reconstructed*	-18.3%

This supports the idea that 3D face recognition across scans with different resolutions can be performed. This aspect is investigated in Sect. IV-E.

C. Reconstruction Under Different Conditions

Results reported in the previous Section have been obtained under constrained conditions in terms of the number of frames processed to extract the super-resolution model, the reference frame used, and the acquisition protocol. In the following, we analyze in more details the effects on the accuracy of the reconstructed model induced by relaxing such conditions.

a) Varying the Number of Frames: In the analysis above, the number of frames used in the reconstruction is kept constant (i.e., $K = 100$ frames per subject are used). Actually, the number of frames used in the reconstruction process affects the accuracy of reconstruction to some extent. Figure 10 reports the error between the high-resolution scan and the model reconstructed using just the first k frames out of K. From the plot, it can be observed as adding just a few frames to the reference one determines an abrupt decrease of the error. This can be motivated by the additional information carried on by the new points included in scans of the face in slightly different positions. This effect is also evident when subjects rotate the head on one side, thus exposing new parts of the face to the sensor (frames 10 to 20); After this point, the error decreases smoothly, with no substantial changes after frame 60. Interestingly, this analysis suggests that a reconstruction with almost the same accuracy of the final super-resolved model can be obtained by using just the first 40/50 frames of the sequence.

b) Varying the Reference Frame: As discussed in Sect. II, the proposed solution performs ICP registration of subsequent frames of a sequence with respect to an initial (reference) frame. In so doing, the first frame defines the projection plane and the initial grid on it. This plane is used to project the

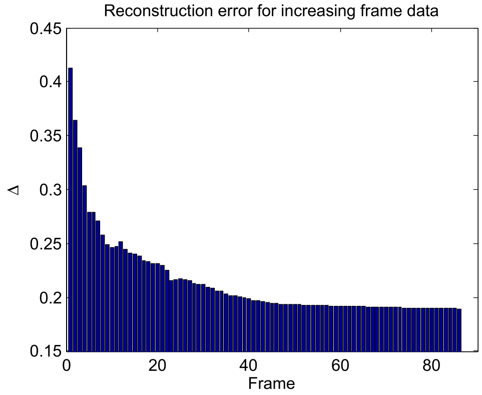


Fig. 10. Mean reconstruction error with respect to the high-resolution scan as a function of the number of frames used in the construction of the super-resolved model.

points of all the subsequent registered frames of the sequence, and to create the super-resolution grid. Due to this, the most important aspect in the selection of the reference frame is its frontal pose. Deviations from this condition can result into erroneous or missing reconstruction of the parts of the face that can be self-occluded by registering to a non-frontal reference pose. In our case, the frontal pose of the initial frame is guaranteed by the acquisition protocol that requires the user to start looking to the camera.

The same result can be achieved by relaxing this constraint, and using instead a solution to automatically detect a frontal frame to be considered as the reference one. For instance, the approach reported in [40] and [41] is capable to estimate the location and orientation of a person's head from a sequence of depth frames, with an average error for the yaw angle of 4° ± 7°, that is always lower than 25°. Accordingly, to investigate the effect that the use of a non-frontal reference frame has on the accuracy of the reconstruction, we computed the super-resolution model using non-frontal reference frames, with a yaw angle deviation up to 25° (i.e., angles of $[-25°, -20°, -15°, -7°, +7°, +15°, +20°, +25°]$ are used). The resulting super-resolution models are reported in Fig. 11. The effect induced by accumulating points of subsequent frames to a non-frontal reference pose can be clearly observed, becoming particularly evident for angles of about $\pm 10°/\pm 15°$. The error distance measure has been also computed between the super-resolution models and the corresponding high-resolution scan, as reported in Fig. 12. Quantitative measures confirm that the approach can cope with yaw deviations from the frontal pose up to about $\pm 7°/\pm 10°$, whereas larger deviations become critical.

c) Acquisition with Facial Expression Changes: The conceived acquisition protocol requires the user to have a neutral and static facial expression during the data acquisition. In this way, data extracted from different frames can be accurately aligned to data of the reference frame. Changes of facial expression during data acquisition may affect the accuracy of reconstruction as the ICP procedure does not cope with elastic deformations. As an example, Fig. 13 shows the error-map of a super-resolution model extracted from data altered by non-rigid deformation of the face during the acquisition process. Figure 13(a) shows some sample RGB frames of a sequence, whereas the super-resolution model and the corresponding error-map are reported in Fig. 13(b). It can be observed the overall accuracy of the reconstruction remains satisfactory, although larger errors emerge in the mouth, eyebrows and eyes regions.

D. Comparative Evaluation

The proposed approach has been compared against two solutions that permit fusion of multiple frames acquired with a Kinect sensor, with the aim to reconstruct an object or scene with a better quality compared to raw data provided by the Kinect sensor: The *Kinect Fusion* approach proposed in [25], which is released as part of the Kinect for Windows SDK; the commercial solution proposed by *Volumental*, which is given as an online service [42] (for the reported experiments, we used the data processing services available through the *Free account*). Both these methods use an acquisition protocol that requires the sensor to be moved around the object (supposed to be fixed) or across the environment to scan. In the proposed application, this protocol is implemented by asking the subject to sit still, and moving the sensor around his/her head at a distance of about 80 to 120*cm*, so as to maintain the best operating conditions for the camera and capture a large view of the face (i.e., the acquired sequence includes the frontal and the left/right side of the face). Compared to the protocol used for constructing super-resolved models, this paradigm is more general, not being constrained to faces, but it also requires substantial human intervention in the acquisition process and an even more constrained scenario, where the subject must remain still.

Since both the methods in [25] and [42] work online on the depth stream produced by the Kinect sensor, we performed an experiment where depth sequences are acquired on-line for four subjects of the UF-S dataset (namely, subjects #009, #014, #016 and #019), and the two methods are used for reconstruction. Figure 14 shows the reconstructed models obtained using the *Kinect Fusion* approach [25], and the corresponding error-maps computed with respect to the high-resolution scans. Compared to the super-resolution models obtained with our approach for the same subjects (see Fig. 9(b) and (d)), a general lower definition of face details can be observed. Results for the same subjects and for the *Volumental* approach [25] are reported in Fig. 15. The main facial traits (i.e., nose, eyebrows, chin) are reasonably defined in the reconstructed models, though finer details are roughly sketched, especially in the mouth and eyes regions.

Using the error measure defined in Sect. IV-B, we also evaluated quantitatively the distance between the models reconstructed with the *Kinect Fusion* and the *Volumental* approaches, and the corresponding high-resolution scans. Results are reported in Table III, and compared with those obtained by our approach. It can be observed, the proposed approach scores the lowest error value.

E. Face Identification Accuracy

One of the goals of this paper is to demonstrate that the use of super-resolution models enables more accurate face

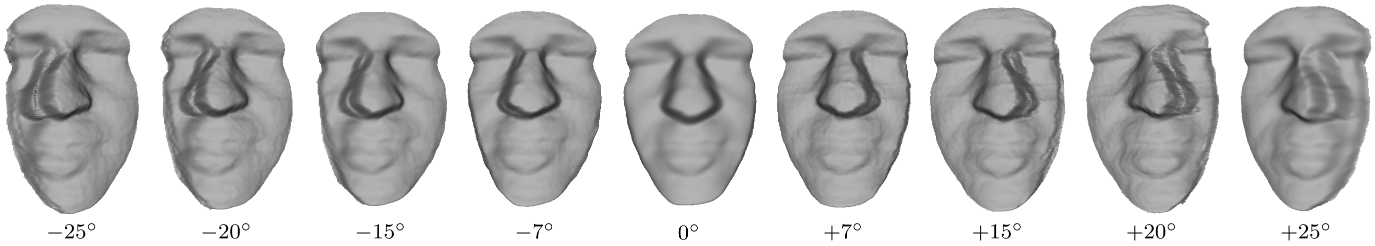


Fig. 11. Face models of a given subject reconstructed with respect to reference frames with a deviation of the yaw angle from the frontal pose (reference angle equal to 0).

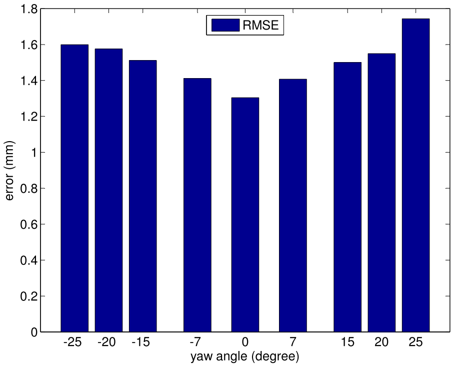


Fig. 12. Distance measure between super-resolution models and the corresponding high-resolution scan as a function of the angular deviation of the pose of the reference frame with respect to the frontal one.

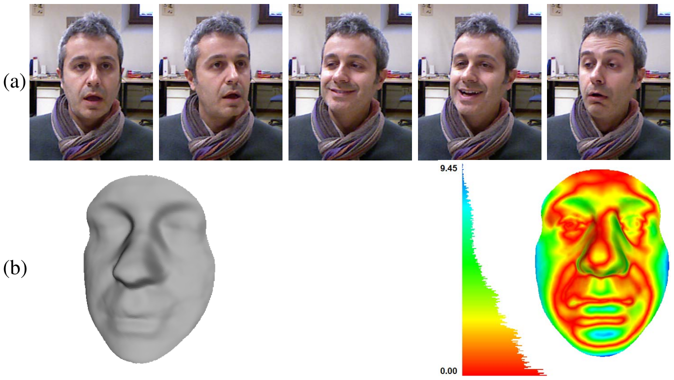


Fig. 13. (a) Sample RGB frames of a sequence where the subject speaks and shows non-neutral facial expressions. (b) The super-resolution model and the corresponding error-map with respect to the high-resolution scan.

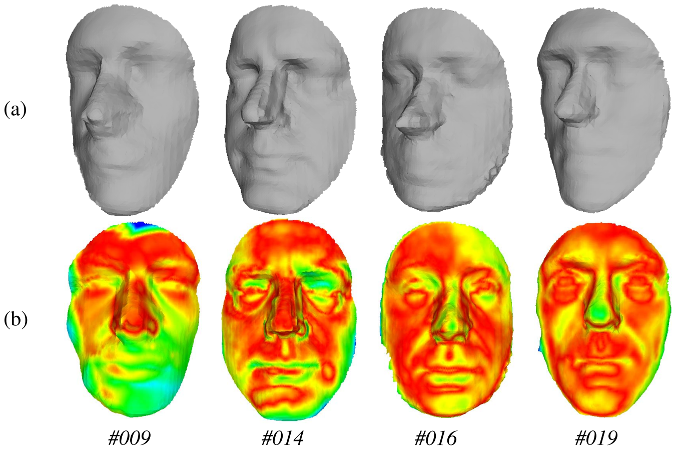


Fig. 14. *Kinect Fusion* [25]: (a) Reconstructed 3D models; (b) error-maps with respect to the high-resolution scans.

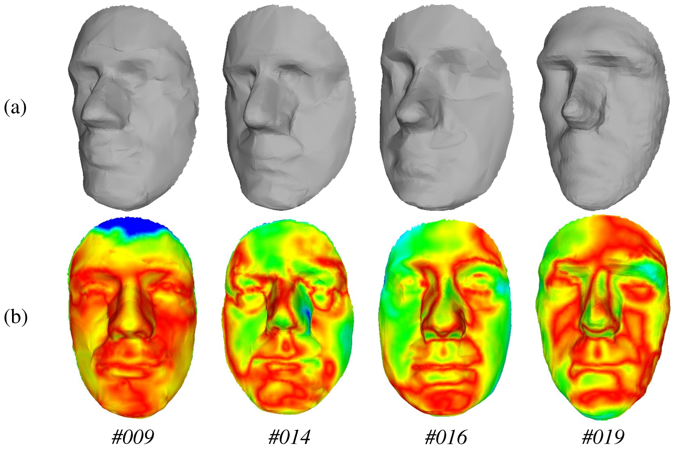


Fig. 15. *Volumental* [42]: (a) Reconstructed 3D models; (b) error-maps with respect to the high-resolution scans.

recognition compared to the use of low-resolution scans. For this purpose, we consider a subject identification task in which the gallery is composed of high-resolution scans, whereas super-resolution models and low-resolution reference scans are used as probes. Description and matching of gallery and probe models is obtained according to the face recognition approach proposed in [7], whose main features are as follows:

- The approach is based on the extraction and comparison of local features of the face. First, SIFT keypoints are detected from a depth image of the face and a subset of them is retained by applying a hierarchical clustering. In this way, a cluster of keypoints with similar position and SIFT descriptors is substituted by a "representative keypoint", thus reducing the overall number of keypoints. Then, the relational information between representative keypoints is captured by measuring how the face depth changes along the surface path connecting pairs of keypoints. The depth values along the path represent a *facial curve* and can be regarded as the result of sectioning the face by a plane passing from the two keypoints and orthogonal to the surface. Face similarity is evaluated by finding correspondences between keypoints of probe and gallery scans, and matching the facial curves

TABLE III

AVERAGE DISTANCE MEASURE COMPUTED BETWEEN THE 3D HIGH-RESOLUTION SCANS AND THE RECONSTRUCTED MODELS OBTAINED, RESPECTIVELY, WITH THE *Kinect Fusion*, *Volumental* AND THE SUPER-RESOLUTION METHOD PROPOSED IN THIS WORK

reconstructed vs. high-res	average distance $RMSE$
Kinect Fusion [25]	1.11
Volumental [42]	1.16
This work	0.84

across the inlier pairs of matching keypoints. In doing so, outlier keypoint correspondences are removed using RANSAC. A statistical model is also used to associate facial curves of the gallery scans with a saliency measure, so that curves that model characterizing traits of some subjects are distinguished from curves that are frequently observed in the face of many different subjects. The approach revealed good performance across different datasets and also in the case of partial face matching. This provides the 3D face recognition approach with the required robustness to manage our scenario.

With this approach, we represented and compared the 3D models used in the identification experiment. We included the 50 high-resolution scans of the UF-S v2.0 subjects in the gallery, and considered the reconstructed models and the reference frames as probes. Recognition accuracy is evaluated through the Cumulative Matching Characteristic (CMC) curves. The accuracy of recognition obtained using the reference frames as probe set is the baseline used for comparison. Figure 16 reports the CMC curves in the case the reference frames (baseline) or the super-resolution models are used as probes. The curves clearly show that super-resolution models achieve a much higher recognition accuracy than raw frames, improving the rank-1 recognition rate from about 58% to 88%. This latter value is close to that obtained when the comparison between different instances of high-resolution scans is considered (represented by the green dashed line in the plot).

For the same gallery and probes, we also computed the average intra-subject distances and the minimum inter-subject distances as reported in Table IV. In this Table, the distances are computed with the measure defined in the 3D face recognition approach that we used in the identification experiment. The Table reports the mean and standard deviation of the intra-subject distance values as well as the minimum inter-subject distance value computed for high-res vs. high-res models, low-res vs. high-res models and super-res vs. high-res models, respectively. It can be noticed that for both the high-resolution and super-resolution models, the mean intra-subject distance value is lower than the minimum inter-subject distance value, whereas this is not the case for the low-resolution models. Compared to the use of low-resolution models, the use of the super-resolution models decreases the mean intra-subject distance yet preserving almost invariate the value of the minimum inter-subject distance. This increase of the gap between mean intra-subject and minimum inter-subject distance values results into higher recognition accuracy.

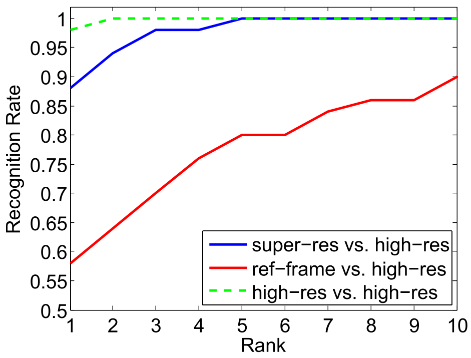


Fig. 16. CMC curves obtained by applying the method in [7] to the super-resolved probes and the reference frame probes (this latter is used as baseline for comparison). Gallery scans are the high-resolution scans. The plot also reports the case in which different instances of high-resolution scans are used as probes (dashed line).

TABLE IV

DISTANCE STATISTICS OBTAINED USING THE FACE RECOGNITION APPROACH IN [7]. DISTANCES BETWEEN HIGH-RESOLUTION SCANS AND, RESPECTIVELY, THE HIGH-RESOLUTION SCANS, THE SUPER-RESOLUTION MODELS AND THE LOW-RESOLUTION REFERENCE FRAMES OF EACH SUBJECT ARE REPORTED

	avg. intra-subject		min. inter-subject
	dist	std-dev	dist
high-res vs. *high-res*	0.003	±0.002	0.0052
low-res vs. *high-res*	0.014	±0.005	0.0100
super-res vs. *high-res*	0.009	±0.003	0.0098

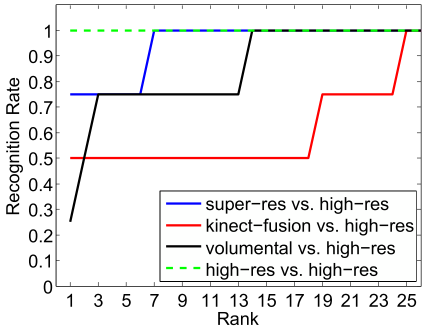


Fig. 17. CMC curves obtained by applying the method in [7] to the models reconstructed with the proposed super-resolution approach, and with the *Kinect Fusion* and *Volumental* methods. Gallery scans are the high-resolution scans. The dashed line also reports the case in which high-resolution scans are used as probes.

Finally, using the same gallery as above, we performed an identification experiment where the probes are the models reconstructed with the proposed method, and with the *Kinect Fusion* and *Volumental* approaches (see Sect. IV-D). Results shown in Fig. 17, confirm the proposed method is capable of achieving better performance than the *Kinect Fusion* and *Volumental*.

V. Discussion and Conclusions

In this paper, we have defined a super-resolution 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 particular, values of the points of the super-resolution model are constructed by iteratively aligning the low-resolution 3D frames to a reference frame (i.e., the first frame of the sequence) using the scaled ICP algorithm, and estimating an approximation function on the cumulated point cloud using Box-spline functions. Qualitative and quantitative experiments have been performed on the released Florence Superface dataset v2.0 that includes, for each subject, a sequence of low-resolution 3D frames and one high-resolution 3D scan used as the ground truth data of a subject's face. In this way, results of the super-resolution process are evaluated by measuring the distance error between the super-resolved models and the ground truth, and by performing face recognition experiments using the super-resolved models as probes and the high-resolution scans as gallery. Results support the idea that constructing super-resolved models from consumer depth cameras can be a viable approach to make such devices deployable in real application contexts that also include identity recognition using 3D faces.

Acknowledgment

The authors acknowledge Prof. Costanza Conti for her valuable comments to the design of the Box-splines based approximation model.

References

[1] Y. Wang, J. Liu, and X. Tang, "Robust 3D face recognition by local shape difference boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 1858–1870, Oct. 2010.

[2] S. Berretti, A. Del Bimbo, and P. Pala, "3D face recognition using isogeodesic stripes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2162–2177, Dec. 2010.

[3] L. J. Spreeuwers, "Fast and accurate 3D face recognition using registration to an intrinsic coordinate system and fusion of multiple region classifiers," *Int. J. Comput. Vis.*, vol. 93, no. 3, pp. 389–414, Mar. 2011.

[4] A. Colombo, C. Cusano, and R. Schettini, "Gappy PCA classification for occlusion tolerant 3D face detection," *J. Math. Imag. Vis.*, vol. 35, no. 3, pp. 193–207, Nov. 2009.

[5] H. Drira, B. B. Amor, A. Srivastava, M. Daoudi, 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.

[6] 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.

[7] S. Berretti, A. Del Bimbo, and P. Pala, "Sparse matching of salient facial curves for recognition of 3D faces with missing parts," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 2, pp. 374–389, Feb. 2013.

[8] A. Savran *et al.*, "Bosphorus database for 3D face analysis," in *Proc. 1st Eur. Workshop Biometrics Identity Manag.*, May 2008, pp. 47–56.

[9] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The Florence 2D/3D hybrid face dataset," in *Proc. Joint ACM Workshop Human Gesture Behavior Understand. (J-HGBU)*, Dec. 2011, pp. 79–80.

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

[11] M. P. 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 (CVPRW)*, Portland, OR, USA, Jun. 2013, pp. 64–69.

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

[13] 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 Workshop Appl. Comput. Vis. (WACV)*, Clearwater, FL, USA, Jan. 2013, pp. 186–192.

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

[15] T. Huang and R. Tsai, "Multi-frame image restoration and registration," *Adv. Comput. Vis. Image Process.*, vol. 1, no. 10, pp. 317–339, 1984.

[16] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1621–1633, Dec. 1997.

[17] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[18] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[19] M. Ebrahimi and E. Vrscay, "Multi-frame super-resolution with no explicit motion estimation," in *Proc. Int. Conf. Image Process., Comput. Vis., Pattern Recognit. (IPCV)*, Las Vegas, NV, USA, Jul. 2008, pp. 455–459.

[20] V. N. Smelyanskiy, P. Cheeseman, D. A. Maluf, and R. D. Morris, "Bayesian super-resolved surface reconstruction from images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2000, pp. 375–382.

[21] S. Peng, G. Pan, and Z. Wu, "Learning-based super-resolution of 3D face model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Genoa, Italy, Sep. 2005, pp. 382–385.

[22] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "LidarBoost: Depth superresolution for ToF 3D shape scanning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 343–350.

[23] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[24] 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.

[25] 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.

[26] 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.

[27] 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, Korea, Aug. 2013, pp. 1–6.

[28] S. Izadi *et al.*, "KinectFusion: Realtime dynamic 3D surface reconstruction and interaction," in *Proc. ACM SIGGRAPH*, Vancouver, BC, Canada, Aug. 2011, p. 1.

[29] S. Berretti, A. Del Bimbo, and P. Pala, "Superfaces: A super-resolution model for 3D faces," in *Proc. 5th Workshops Non-Rigid Shape Anal. Deformable Image Alignment (NORDIA)*, Firenze, Italy, Oct. 2012, pp. 73–82.

[30] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, Sep. 1987.

[31] S. Du, N. Zheng, L. Xiong, S. Ying, and J. Xue, "Scaling iterative closest point algorithm for registration of m-D point sets," *J. Vis. Commun. Image Represent.*, vol. 21, nos. 5–6, pp. 442–452, Jul. 2010.

[32] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing. I. Theory," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 821–833, Feb. 1993.

[33] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing. II. Efficiency design and applications," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 834–848, Feb. 1993.

[34] M. Charina, C. Conti, K. Jetter, and G. Zimmermann, "Scalar multivariate subdivision schemes and box splines," *Comput. Aided Geometric Design*, vol. 28, no. 5, pp. 285–306, Jun. 2011.

[35] T. Huynh, R. Min, and J.-L. Dugelay, "An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data," in *Proc. Comput. Vis.-ACCV Workshops*, Daejeon, Korea, Nov. 2012.

[36] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proc. 6th IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Arlington, VA, USA, Sep. 2013, pp. 1–6.

[37] (2010). *3dMD*. [Online]. Available: http://www.3dmd.com

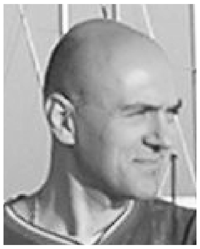
[38] C. Xu, T. Tan, Y. Wang, and L. Quan, "Combining local features for robust nose location in 3D facial data," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1487–1494, Oct. 2006.

[39] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proc. 3rd Int. Conf. 3-D Digit. Imag. Model. (3DIM)*, May 2001, pp. 145–152.

[40] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. German Assoc. Pattern Recognit. (DAGM) Symp.*, Frankfurt, Germany, Aug. 2011, pp. 1–10.

[41] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[42] Volumental. (2013). *3D Scanning Solutions*. [Online]. Available: http://www.volumental.com/



Stefano Berretti received the Laurea degree in electronic engineering and the Ph.D. degree in informatics engineering and telecommunications from the University of Florence, Florence, Italy, in 1997 and 2001, respectively. He is currently an Associate Professor with the Department of Information Engineering and the 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. His recent research has addressed 3D object retrieval and partitioning, 3D/4D face, and facial expression recognition. On this latter subject, in 2009, he was a Visiting Professor with the Institute TELECOM, TELECOM Lille 1, Lille, France. In 2013, he was also a Visiting Professor with the Khalifa University of Science Technology and Research, Sharjah, United Arab Emirates.

Prof. Berretti was the Co-Chair of the 2012 Fifth Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment, in conjunction with ECCV 2012. He is also a frequent reviewer of several journals in the area of multimedia and pattern recognition, and of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

Pietro Pala received the Laurea degree in electronic engineering and the Ph.D. degree in information and telecommunications engineering from the University of Florence, Florence, Italy, in 1994 and 1997, respectively. He is currently an Associate Professor with the University of Florence, where he teaches Image and Video Analysis and Database Management Systems with the School of Information Engineering and Fundamentals of Multimedia and Programming Languages at the Master in Multimedia Content Design. His research activity has focused on the use of pattern recognition models for multimedia information retrieval and biometrics. Former studies targeted the definition of elastic models for measuring shape similarity and support shape-based retrieval in image databases. From these studies, a number of different yet related topics were investigated, including image segmentation, content-based description and retrieval of color images, multidimensional indexing structures for retrieval by color and shape, semantic content description in paintings and advertising videos, description and similarity matching of 3D models, and segmentation of 3D models. Recently, the research activity focuses on the study of biometric models for person recognition based on 3D facial scans.

Alberto del Bimbo is a Full Professor of Computer Engineering, the Director of the Master in Multimedia, and the Director of the Media Integration and Communication Center with the University of Florence, Florence, Italy. He was the Deputy Rector of 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 published over 250 publications in some of the most distinguished scientific journals and international conferences, and has authored the monography, *Visual Information Retrieval*. From 1996 to 2000, he was the President of the IAPR Italian Chapter, and from 1998 to 2000, a Member-at-Large of the IEEE Publication Board. He was the General Chair of IAPR ICIAP'97, the International Conference on Image Analysis and Processing, the IEEE ICMCS'99, the International Conference on Multimedia Computing and Systems, and the Program Co-Chair of ACM Multimedia 2008. He was also the General Co-Chair of ACM Multimedia 2010 and the European Conference on Computer Vision in 2012. He is an IAPR Fellow and an Associate Editor of *Multimedia Tools and Applications*, *Pattern Analysis and Applications*, *Journal of Visual Languages and Computing*, and *International Journal of Image and Video Processing*, and was an Associate Editor of *Pattern Recognition*, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.