Increasing 3D Resolution of Kinect Faces

Stefano Berretti, Pietro Pala, and Alberto del Bimbo

University of Florence, Italy

Abstract. Performing face recognition across 3D scans of 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. However, these devices have still a much lower resolution than the 3D high-resolution scanners typically used for face recognition applications. Due to this, comparing low- and high-resolution scans can be misleading. Based on these considerations, in this paper we define an approach for reconstructing a higher-resolution 3D face model from a sequence of low-resolution 3D scans. The proposed solution uses the scaled ICP algorithm to align the low-resolution scans with each other, and estimates the value of the high-resolution 3D model through a 2D Box-spline approximation. The approach is evaluated on the *The Florence face* dataset that collects high- and low-resolution data for about 50 subjects. Measures of the quality of the reconstructed models with respect to high-resolution scans and in comparison with two alternative techniques, demonstrate the viability of the proposed solution.

Keywords: Kinect camera; 3D super-resolution; 2D box-splines

1 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 in terms of gender, age, ethnicity, occlusions due to hair or external accessories. The resolution at which 3D face scans are acquired changes across different datasets, but it is typically the same within one dataset. 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 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 at about 30fps, with a resolution of 18ppi at a distance of about 80cm from the

sensor. 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], continuous authentication [4] and recognition [5–7] 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 expected to be of lower resolution and potentially acquired with 4D cameras.

In this context, reducing the impact on the recognition accuracy due to the match of low-resolution probes against high-resolution gallery scans is relevant, 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, this could open the way to more versatile 3D face recognition methods deployable in contexts where the acquisition of high resolution 3D scans is not convenient or even possible. Based on these premises, in this work we aim to provide an effective approach specifically tailored to reconstruct a higher-resolution face model from a sequence of low-resolution depth frames, thus capable of reducing the gap between low- and high-resolution acquisitions.

1.1 Related work

Methods to recover one high-resolution image from a set of low-resolution images possibly altered by noise, blurring or geometric warping, have been formerly introduced for 2D still images [8–12], and go under the term of super-resolution.

Super-resolution techniques have been also applied to 3D generic data [13, 14]. Previous work that focus in particular on super-resolution of 3D faces are reported in [15, 16]. In [15], high resolution 3D face models are used 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 [16] the super-resolution process is modeled as a progressive resolution chain, whose features are computed as the solution to a MAP problem. However, in both the cases, the framework is validated just on synthetic data.

Methods in [17, 18] and [19] approach the problem of noise reduction in depth data by fusing the observations of multiple scans to construct one denoised scan. 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 object. Later, dynamic interaction has been added to the system in [20], 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 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], 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 [19] 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 (NF), 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. In [21], a method to increase the resolution of the face scans acquired with a Kinect is proposed. The method is based on ICP registration on the first frame of the sequence and subsequent points approximation, but results are quite preliminary and no evidence that the approach is indeed capable of producing a super-resolution is provided.

1.2 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 Kinect camera. In the proposed approach, first, the region containing the face is automatically detected and cropped in each depth frame; then, the face of the first frame is used as reference and all the faces from the other frames are aligned to the reference; finally, the aggregated data of these multiple aligned observations are resampled at a higher resolution and approximated using 2D-Box splines. The proposed approach has been evaluated on the *The Florence face* dataset, which includes, for each individual, one Kinect depth sequence and one high-resolution face scan acquired through a 3dMD scanner. 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, with the proof the proposed approach is capable of producing a super-resolved 3D model rather than just a denoised one;
- An evaluation demonstrating the accuracy of the reconstructed super-resolved models with respect to the high-resolution scans, and in comparison to two alternative solutions.

The rest of the paper is organized as follows: The problem statement and the basic notation are defined in Sect. 2; The super-resolution approach based on facial data approximation is described and validated in Sect. 3. Experimental results are reported and discussed in Sect. 4. Finally, discussion and conclusions are given in Sect. 5.

2 Problem statement

In this work, we aim to reconstruct 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). In particular, low-resolution frames are acquired by a Kinect camera placed in front of a sitting subject, while s/he is slightly rotating the head to the left and right side. In Fig. 1(a), a sample depth frame is shown. The face region is cropped in each frame by using the *Face Tracking* function available in the device SDK, as shown in Fig. 1(b).

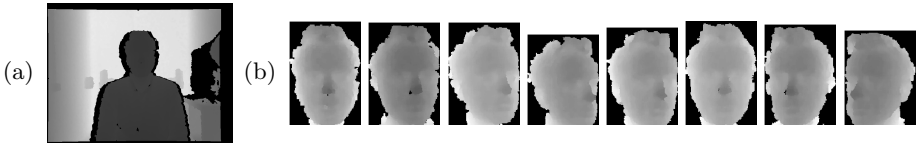

Fig. 1. (a) Sample depth frame acquired by the Kinect; (b) Some cropped frames from the sequence, with the pose of the face varying from frontal to right and left side.

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 $\{X_L^{(k)}\}$ 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 mainly for the geometric transformation of X_H to the coordinates of the k-th low-resolution frame $X_L^{(k)}$, the blurring effect induced by the atmosphere and camera lens, down-sampling, and additive noise. In particular, we note 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*. The geometric transformation that maps the coordinate systems of subsequent low-resolution frames to the first frame of the sequence is computed with a variant of the ICP algorithm, which jointly estimates the 3D rotation and translation parameters as well as the scaling one [22] (this operation is applied just to the cropped region of the face). The data cumulated by this process represent a cloud of points in the 3D space, and these points are regarded as observations of the value of the high-resolution image X_H.

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 translation, rotation and scaling transformation that best aligns the data:

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

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 [22].

The ICP algorithm usually requires an appropriate initialization to avoid convergence to local minima. For this purpose, alignment of the generic frame $X_L^{(k)}$ to 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 then used for fine registration.

3 Increasing the face resolution

Based on the procedure described so far, data 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 set of all these scattered data points $\left\{\mathbf{P}^{(j)}\right\}_{j=1}^{J} = \left\{(P_x^{(j)}, P_y^{(j)}, P_z^{(j)})\right\}_{j=1}^{J}$ represent the observed samples of the underlying face surface, which is approximated through a function $\Gamma(x,y)$. This function is defined on a high resolution uniform grid Φ compared to the low resolution uniform grid Ω of the reference frame $X_L^{(1)}$. It should be noticed that, under the effect of Eq. (2), data points are scattered and distributed irregularly with respect to both the high and low resolution grids Φ and Ω. The approximation model acts as a function $\Gamma(x,y)$ that given the set of scattered points $\left\{\mathbf{P}^{(j)}\right\}_{j=1}^{J}$ that are expected to sample the 2D facial surface in the 3D space, projects them onto a reference plane Π (the (x,y) plane of the first frame) and then estimates the *height* of the surface for a generic point $p \in \Pi$ within the convex hull of the projected set of points (see Fig. 2). In this way, given the super-resolution uniformly spaced grid Φ in Π, it is possible to estimate the value of the 2D facial surface for each point of Φ enclosed within the convex hull of the projection of the scattered points onto Π.

To estimate the approximating function, the 2D Box-splines model is used [23]. Accordingly, the approximating function $\Gamma(x,y)$ is expressed as a weighted sum of Box splines originated by translation of a 2D base function $B_{0,0}(x,y)$ with local support. Given a 1D lattice $\{x_{-n},\ldots,x_{-1},x_0,x_1,\ldots,x_n\}$, the 1D first degree ($C^0$ continuity) base function $b_0(t)$ is defined as:

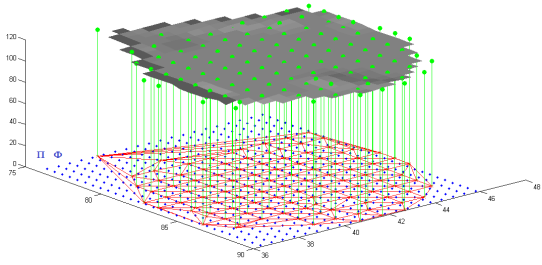

Fig. 2. The projection of points of frames in a sequence onto the reference plane associated to the first frame distribute irregularly. Estimation of values of the underlying surface (shown in gray) on a regular grid (blue points) is obtained by computing one approximating function that fits the data.

$$b_0(t) = \begin{cases} 0 & if\ t \in (-\infty, x_{-1}] \\ \frac{t-x_{-1}}{x_0-x_{-1}} & if\ t \in (x_{-1}, x_0] \\ \frac{x_1-t}{x_1-x_0} & if\ t \in (x_0, x_1] \\ 0 & if\ t \in (x_1, \infty)\ . \end{cases} \tag{3}$$

The translated 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 possible by considering a 2D lattice $\{x_{i,j}\}$ and the 2D base function $B_{0,0}(x, y)$ computed as the tensor product of the 1D base function:

$$B_{0,0}(x, y) = b_0(x)b_0(y)\ . \tag{4}$$

The translated copy of the base function, centered on the generic node $x_{i,j}$ 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 and with local support, being zero for all points (x, y) not included in any of the rectangular cells with one vertex on $x_{i,j}$. The function $\Gamma(x, y)$ is expressed as:

$$\Gamma(x, y) = \sum_{i,j} w_{i,j} B_{i,j}(x, y)\ , \tag{5}$$

being $w_{i,j}$ the set of weights that yield the best approximation to the points 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 data 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 data points. This constraint is expressed by K equations of the form:

$$\Gamma(P_x^{(k)}, P_y^{(k)}) = P_z^{(k)} \quad k = 1, \ldots, K\ . \tag{6}$$

Due to the form of the basis functions (Eqs. (3)-(4)), $\Gamma(x, y)$ is continuous everywhere. Since $\Gamma(x, y)$ is not derivable in correspondence to the points of the lattice $\{x_{i,j}\}$, its smoothness is forced by the following set of equations:

$$\begin{aligned} \left.\frac{\partial^+ \Gamma(x,y)}{\partial x}\right|_{x_{ij}} &= \left.\frac{\partial^- \Gamma(x,y)}{\partial x}\right|_{x_{ij}} \\ \left.\frac{\partial^+ \Gamma(x,y)}{\partial y}\right|_{x_{ij}} &= \left.\frac{\partial^- \Gamma(x,y)}{\partial y}\right|_{x_{ij}} \qquad i,j = -n,\ldots,n\ . \end{aligned} \tag{7}$$

The left and right partial derivatives of Eq. (7) can be obtained analytically, and combined with Eq. (6) represent a system of $K + n^2$ linear equations in the n^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.

3.1 Resolution gain

The proposed solution results in a face surface with an increased resolution, rather than just in a surface denoising. This can be shown considering the reference frame of a sample sequence in Fig. 3(a), and the reconstruction obtained from the depth sequence of the same face at different resolutions, namely, 104×157, 207×313 and 413×625, as reported in Fig. 3(b)-(d), respectively.

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 $M = \zeta N$, and $\mathcal{R}(\cdot)$ 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:

$$\eta(\zeta) = \frac{\sum_{i,j} |\mathcal{R}(F_{\zeta-1}) - F_\zeta|}{\zeta^2 N^2}\ . \tag{8}$$

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

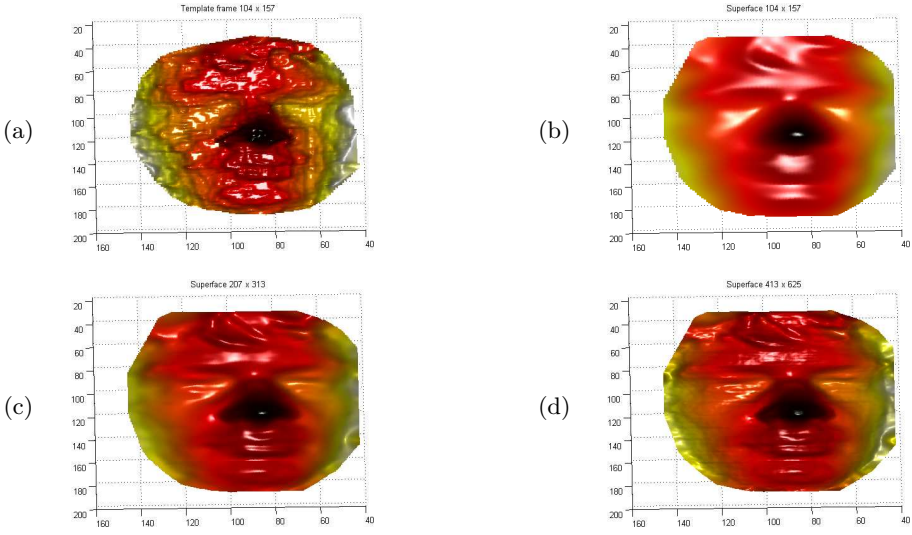

Fig. 3. (a) Reference frame of a sequence; (b)-(d) Three models reconstructed at resolutions, respectively, 104×157 (same resolution as the original, just denoising), 207×313, and 413×625.

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. 4, 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.

4 Experimental results

The proposed approach has been evaluated considering the accuracy of the super-resolution reconstruction, by computing the error between the super-resolved models and the corresponding high-resolution scans (Sect. 4.1). In so doing, we also compared our approach against two alternative solutions (Sect. 4.2).

The study reported hereafter has been performed on the *The Florence face* dataset (UF-S) [24]. Some public datasets exist for face analysis from consumer cameras, like Kinect (see for example the EURECOM Kinect Face dataset [25], or the The 3D Mask Attack database specifically targeted to detect face spoofing attacks [26]). However, to the best of our knowledge the UF-S dataset is the only one providing sequences of low resolution face scans acquired with the Kinect

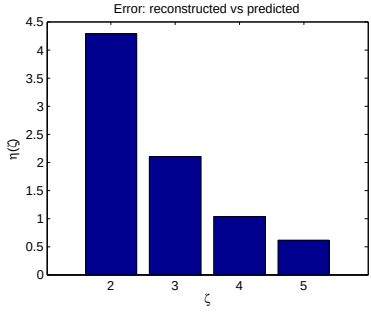

Fig. 4. Values of $\eta(\zeta)$ measure the error between the model reconstructed through the proposed super-resolution approach at the resolution gain ζ, and the prediction (by bilinear interpolation) based on the model reconstructed at the resolution gain ζ-1.

camera and high resolution 3D scans, for the same subjects. This dataset enrolls 50 subjects, each with the following data:

- A 3D high-resolution face scan, with about 40,000 vertices. 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;
- A video sequence acquired with the *Kinect* camera. During acquisition the person sits in front of the sensor at an approximate distance of 80cm. The subject is also 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. This results in video sequences lasting approximately 10 to 15 seconds on average, at 30fps.

Fig. 5. Sample of the *The Florence face* dataset: (a) 3D high-resolution scan; (b) RGB and depth frames from the Kinect video sequence, with the head pose changing from frontal to left and right side.

The 3D high-resolution scans and the Kinect video sequences are provided in the form produced by the sensors, without any processing or annotation. Figure 5 shows samples of the raw data acquired for a subject (RGB frames of the sequence are also reported, but they are not used in our solution).

4.1 Reconstruction accuracy

The first evaluation aims to show the error between the reconstructed 3D super-resolution model with respect to the 3D high-resolution scan of a same subject, also in comparison to the same measure of error computed 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 small areas of the face are not visible to the sensor due to self-occlusion effects.

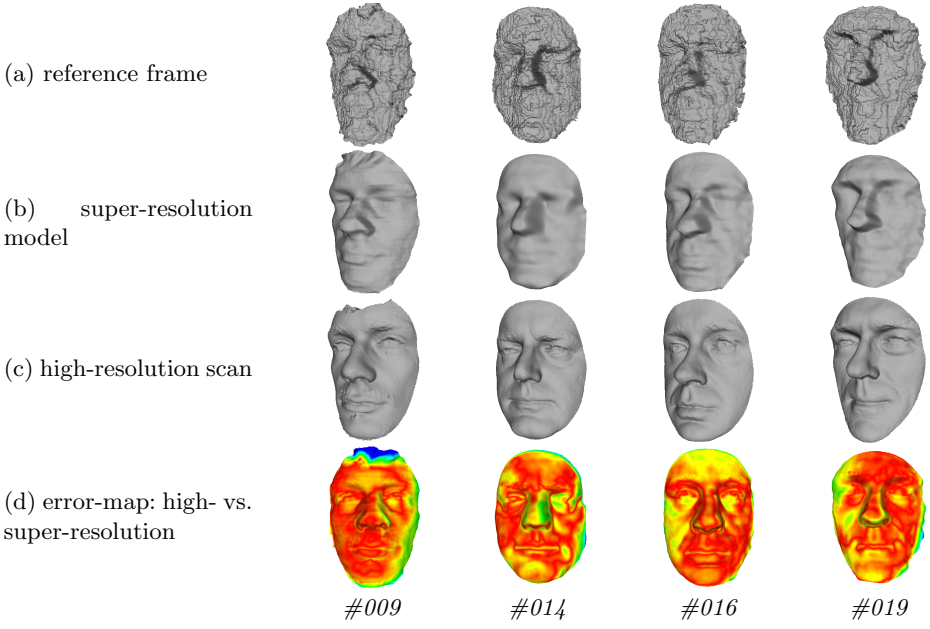

Fig. 6. Each column corresponds to a different subject and reports: (a) The low resolution 3D scan of the reference frame; (b) The super-resolution 3D model; (c) The high-resolution 3D scan. The error-map in (d) shows, for each point of the super-resolution model, the value of the distance to its closest point on the high-resolution scan after alignment (distance increases from red/yellow to green/blue).

All the subjects in the UF-S 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 $95mm$ centered at the nose tip (the approach in [27] is used to detect 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 ICP

registration [28]. Then, for each point of the super-resolution model its distance to the closest point in the high-resolution scan is computed to build an error-map. As an example, Fig. 6 shows for some representative subjects (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).

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 $p' \in S'$:

$$RMSE(S, S') = \left(\frac{1}{N} \sum_{i=1}^{N} (p_i - p'_i)^2 \right)^{1/2}, \tag{9}$$

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

Table 1. The first two rows report the average RMSE between the 3D *high-resolution* scan and, respectively, the *super-resolution* model and the *reference* scan of *same subjects*. In the third row, the average RMSE between any two high-resolution scans of *different subjects* is reported. The rightmost column also evidences the relative variation of the intra-subjects distance values with respect to the inter-subject distance

	models	average $RMSE$	% variation
same subject	*reference* vs. high-res	1.48	+4.2%
	reconstructed vs. high-res	1.16	-18.3%
different subjects	*high-res* vs. high-res	1.42	–

Results obtained using this distance measure are summarized in Table 1. 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 1 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 (more than 20% decrease of the RMSE passing from the first to the second row). To better emphasize the actual improvement, the average inter-subject distance between any two high-resolution scans of different subjects is also reported in the last row of Table 1. The relative variation of the intra-subject distance values in the first two rows compared to the inter-subject high-resolution distance values is reported in the rightmost column in the Table. It

can be noticed that compared to the average inter-subject distance, the accuracy of the super-resolution models is considerable higher than the accuracy of the reference scans. This supports the idea that 3D face recognition across scans with different resolutions can be performed.

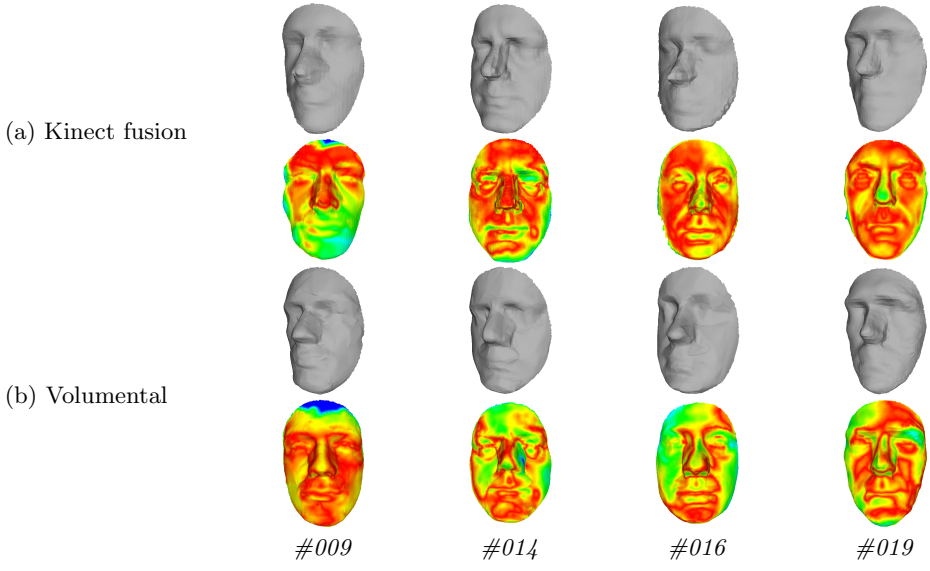

Fig. 7. (a) *Kinect Fusion* [17]: (b) *Volumental* [29]. In both the cases, the reconstructed 3D models and the corresponding error-maps with respect to the high-resolution are reported in the top and bottom row.

4.2 Comparative evaluation

The proposed approach has been compared against two solutions that permit fusion of multiple frames acquired with a Kinect sensor: The *Kinect Fusion* approach proposed in [17], which is released as part of the Kinect for Windows SDK; the commercial solution proposed by *Volumental*, which is given as an online service [29] (for the reported experiments, we used the data processing service 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 120cm, 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.

Figure 7(a) shows the reconstructed models obtained using the *Kinect Fusion* approach [17], 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. 6(b) and (d)), a general lower definition of face details can be observed. Results for the same subjects and for the *Volumental* approach [17] are reported in Fig. 7(b). 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.

Table 2. 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 $RMSE$
Kinect Fusion [17]	1.11
Volumental [29]	1.16
This work	0.84

Using the error measure defined in Sect. 4.1, 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 2, and compared with those obtained by our approach. It can be observed, the proposed approach scores the lowest error value.

5 Discussion and conclusions

In this paper, we have defined an approach that permits the construction of a super-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 *The Florence face* dataset 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. 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.

References

1. Passalis, G., Perakis, P., Theoharis, T., Kakadiaris, I.A.: Using facial symmetry to handle pose variations in real-world 3D face recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **33**(10) (October 2011) 1938–1951
2. Drira, H., Ben Amor, B., Srivastava, A., Daoudi, M., Slama, R.: 3D face recognition under expressions, occlusions, and pose variations. IEEE Trans. on Pattern Analysis and Machine Intelligence **35**(9) (2013) 2270–2283
3. Pamplona Segundo, M., Silva, L., Bellon, O.: Real-time scale-invariant face detection on range images. In: Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC), Anchorage, Alaska, USA (October 2011) 914–919
4. Pamplona Segundo, M., Sarkar, S., Goldgof, D., Silva, L., Bellon, O.: Continuous 3d face authentication using rgb-d cameras. In: Proc. IEEE Work. on Biometrics, Portland, Oregon, USA (June 2013) 1–6
5. Min, R., Choi, J., Medioni, G., Dugelay, J.L.: Real-time 3D face identification from a depth camera. In: Proc. Int. Conf. on Pattern Recognition (ICPR), Tsukuba, Japan (November 2012) 1739–1742
6. Li, B.Y.L., Mian, A.S., Liu, W., Krishna, A.: Using kinect for face recognition under varying poses, expressions, illumination and disguise. In: Proc. IEEE Work. on Applications of Computer Vision (WACV), Clearwater, Florida (January 2013) 186–192
7. Goswami, G., Bharadwaj, S., Vatsa, M., Singh, R.: On RGB-D face recognition using Kinect. In: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS), Washington DC, USA (September 2013)
8. Huang, T., Tsai, R.: Multi-frame image restoration and registration. Advances in Computer Vision and Image Processing **1**(10) (1984) 317–339
9. Hardie, R., Barnard, K., Armstrong, E.: Joint map registration and high-resolution image estimation using a sequence of undersampled images. IEEE Trans. on Image Processing **6**(12) (1997) 1621–1633
10. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. on Pattern Analysis and Machine Intelligence **24**(9) (2002) 1167–1183
11. Farsiu, S., Robinson, M., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Trans. on Image Processing **13**(10) (2004) 1327–1344
12. Ebrahimi, M., Vrscay, E.: Multi-frame super-resolution with no explicit motion estimation. In: Proc. Int. Conf. on Image Processing, Computer Vision, and Pattern Recognition (IPCV), Las Vegas, Nevada, USA (July 2008) 455–459
13. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Minneapolis, Minnesota, USA (June 2007) 1–8
14. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: Lidarboost: Depth superresolution for ToF 3D shape scanning. In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), Miami, Florida, USA (June 2009) 343–350
15. Peng, S., Pan, G., Wu, Z.: Learning-based super-resolution of 3D face model. In: Proc. IEEE Int. Conf on Image Processing (ICIP). Volume II., Genoa, Italy (September 2005) 382–385
16. Pan, G., Han, S., Wu, Z., Wang, Y.: Super-resolution of 3D face. In: Proc. European Conf. on Computer Vision (ECCV), Graz, Austria (May 2006) 389–401
17. Newcombe, R., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: Proc. IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR), Basel, Switzerland (October 2011) 1–10

18. Hernandez, M., Choi, J., Medioni, G.: Laser scan quality 3-d face modeling using a low-cost depth camera. In: Proc. European Signal Processing Conf. (EUSIPCO), Bucharest, Romania (August 2012) 1995–1999
19. Choi, J., Sharma, A., Medioni, G.: Comparing strategies for 3D face recognition from a 3D sensor. In: Proc. IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN), Gyeongju, Korea (August 2013) 1–6
20. Izadi, S., Newcombe, R., Kim, D., Hilliges, O., Molyneaux, D., Hodges, S., Kohli, P., Shotton, J., Davison, A., Fitzgibbon, A.: Kinectfusion: realtime dynamic 3D surface reconstruction and interaction. In: Proc. ACM SIGGRAPH, Vancouver, Canada (August 2011) 1
21. Berretti, S., Del Bimbo, A., Pala, P.: Superfaces: A super-resolution model for 3D faces. In: Proc. Work. on Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA), Firenze, Italy (October 2012) 73–82
22. Du, S., Zheng, N., Xiong, L., Ying, S., Xue, J.: Scaling iterative closest point algorithm for registration of m-D point sets. Journal of Visual Communication and Image Representation **21** (2010) 442–452
23. Charina, M., Conti, C., Jetter, K., Zimmermann, G.: Scalar multivariate subdivision schemes and box splines. Computer Aided Geometric Design **28**(5) (2011) 285–306
24. The Florence face dataset. http://www.micc.unifi.it/datasets/4d-faces/ (2013)
25. Huynh, T., Min, R., Dugelay, J.L.: An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In: ACCV Work. on Computer Vision with Local Binary Pattern Variants, Daejeon, Korea (November 2012)
26. Erdogmus, N., Marcel, S.: Spoofing in 2D face recognition with 3D masks and anti-spoofing with kinect. In: IEEE Int. Conf. on Biometrics: Theory, Applications and Systems, (BTAS), Washington DC, USA (September 2013)
27. Xu, C., Tan, T., Wang, Y., Quan, L.: Combining local features for robust nose location in 3D facial data. Pattern Recognition Letters **27**(13) (2006) 1487–1494
28. Rusinkiewicz, S., Levoy, M.: Efficient variants of the ICP algorithm. In: Proc. Int. Conf. on 3D Digital Imaging and Modeling (3DIM), Quebec City, Canada (May 2001) 145–152
29. Volumental. http://www.volumental.com/ (2013)