Superfaces: A Super-Resolution Model for 3D Faces

Stefano Berretti, Alberto Del Bimbo, and Pietro Pala

University of Firenze, Italy

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 of primal importance after the introduction of a new generation of low cost 4D scanning devices. These devices are capable of combined depth/rgb acquisition over time with a low resolution compared to the 3D scanners typically used in 3D face recognition benchmarks. In this paper, we define a super-resolution model for 3D faces by which a sequence of low-resolution 3D scans can be processed to extract a higher resolution 3D face model, namely the *superface* model. The proposed solution relies on the Scaled ICP procedure to align the low-resolution 3D models with each other and estimate the value of the high-resolution 3D model based on the statistics of values of the low-resolution scans in corresponding points. The approach is validated on a data set that includes, for each subject, one sequence of low-resolution 3D face scans and one ground-truth high-resolution 3D face model acquired through a high-resolution 3D scanner. In this way, results of the super-resolution process are evaluated qualitatively and quantitatively by measuring the error between the superface and the ground-truth.

1 Introduction

In recent years, many approaches have been presented to support person recognition by the analysis of 3D face models. In this research area, many challenging issues have been successfully investigated, including 3D face recognition in the presence of non-neutral facial expressions [1, 2], occlusions [3], and missing data [4], to say a few. Typically, the proposed solutions are tested following well defined evaluation protocols on consolidated benchmark data sets that, in order to obtain a reasonable coverage of the many different traits and characteristics of the human face, include 3D face models from several persons differing in terms of gender, age, ethnicity, hair style and accessories (spectacles, nose rings, etc.). The resolution of the 3D face models changes across different data sets, but is always the same within one data set. The issues related to the resolution of the 3D face model and its impact on the recognition accuracy have not been addressed explicitly in the past. Nevertheless, the relevance of these issues is increasing, motivated by the introduction in the marketplace of a new generation of low cost 4D scanning devices (such as Microsoft® Kinect or Asus® Xtion PRO LIVE)

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 73–82, 2012.

that are capable of combined depth/rgb acquisition over time (30 fps) with a resolution of 18 ppi at a distance of about 30 inches from the scanning device. Evaluating the impact on the recognition accuracy of matching one low-res probe to a high-res gallery is certainly one issue, but an even more challenging issue addresses the study of models to reconstruct one super-resolution face image out of the many low-res depth frames acquired by the 4D scanner.

Formerly introduced for images, super-resolution is the process that aims at recovering one high-resolution image from a set of low-resolution images possibly altered by noise, blurring or geometric warping [5–9]. Approaches proposed in the literature that use super-resolution models in the specific context of 3D data can be grouped in two distinct classes: approaches that apply the super-resolution in the 2D space and then use multiple super-resoluted 2D image to reconstruct a super-resolution 3D object [10]; and approaches that operate directly in the 3D space by applying the super-resolution model on 3D data [11–14]. The approach proposed in [13] is conceived to operate on data provided by time-of-flight cameras. These 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. Also the approach proposed in [12] targets processing of data provided by time-of-flight cameras. However, the proposed 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 and a regularization term to cope with reconstruction of sparse data points. In general, the 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. This aspect is less relevant if the 3D data represent a single object with smooth surface such as a face. The approaches proposed in [11, 14] address the specific problem of super-resolution of facial models. In [11], 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 [14] the super-resolution process is modeled as a progressive resolution chain whose features are computed as the solution to a MAP problem.

In this paper we present a model to derive one super-resolution 3D face from several low-res depth images acquired through a Microsoft® Kinect scanner. The proposed approach develops on the super-resolution model proposed in [9] and combines three main processing modules, namely the *face detector*, the *face registration* and the *face sampler*. The face detector processes each frame acquired by the 4D scanner so as to detect and crop the region of the frame where the face is represented. These cropped faces are used to feed the face registration module that performs 3D alignment of all the cropped faces to the first one, used as template. In this way a layered representation is built which provides,

for each point on the template several observation values. Based on the statistics of these observation values, the face sampler module resamples the data at a higher resolution.

To validate the proposed approach and estimate the accuracy of the computed superface models we set up a data set of heterogeneous face models, described in detail in Sect. 3.1, that includes, for each individual, one sequence of depth images acquired through a Microsoft® Kinect scanner as well as one high-resolution face model acquired through a 3dMD® scanner. In this way, the accuracy of the reconstructed superface model can be quantitatively measured by comparing the reconstructed model to the corresponding high-res model.

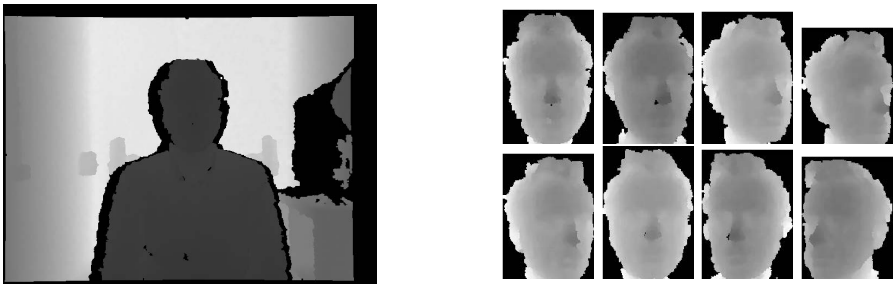
Hence, the contribution of this paper is twofold: we describe a model to extract one super-resolution 3D face model out of a sequence of several low-res depth facial images; we set up and give public access to a data set of heterogeneous face models to be used by researchers working on this topic. The paper is organized as follows: Problem statement and the adopted notation are defined in Sect. 2. The description of the modules for the detection of the facial region in the acquired depth frames, for pairwise alignment of facial data across different frames, and for resampling of facial data are described in Sect. 2.1, Sect. 2.2 and Sect. 2.3, respectively. Finally experimental results and conclusions are discussed in Sect. 3 and Sect. 4.

2 The Superface Model

In the literature, the super-resolution process is typically formalized 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-res images represent new samples of the patches in the observed scene. No improvement on resolution (if any, only in terms of SNR) would be possible from images deriving from a fixed camera observing a static scene. Let $\Omega = [1, \ldots, N] \times [1, \ldots, M]$ and $\Phi = [1, \ldots, zN] \times [1, \ldots, zM]$ be the sampling grids of the low and high resolution images, being z a positive integer representing the resolution gain. The forward degradation model, describing the formation of the low-res images 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-res images, X_H the high-res image and P_k the operator that maps the high-res image onto the coordinate system and sampling grid of the k-th low-res image. 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-res image $X_L^{(k)}$; *ii*) blurring introduced by the effect of the atmosphere and camera lens; *iii*) downsampling and *iv*) additive noise.



(a) (b)

Fig. 1. (a) One sample frame (depth image) acquired by the scanning device. (b) Some cropped faces from the sequence of acquired frames.

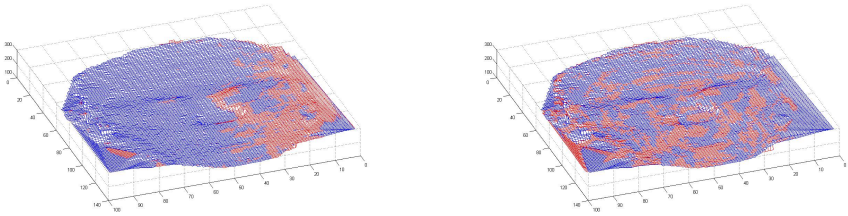
The coordinate system of the high-res image X_H is aligned to the coordinate system of the first low-res image $X_L^{(1)}$. Computation of the geometric transformation that maps the coordinate systems of subsequent low-res images is operated by registration of the low-res images. This is accomplished using the Iterative Closest Point procedure, as described in Sect. 2.2.

2.1 Face Cropper

Low-res images correspond to frames (depth images) acquired by a Microsoft® Kinect scanner placed in front of a subject standing at a distance of approx 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) one sample frame out of the sequence of depth images acquired by the scanner is shown. 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).

2.2 Face Registration

As anticipated before, computation of the geometric transformation that aligns low-res images to a common reference system is accomplished through a variant of the base Iterative Closest Point procedure [15] that jointly estimates the 3D rotation and translation parameters as well as the scaling one [16]. 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 accomplished by computing the similarity transform (translation, rotation and scaling) that best aligns the transformed data to the



(a) (b)

Fig. 2. Facial data acquired in two sample frames (one with red and one with blue colors) before (a) and after (b) the application of the adopted ICP procedure

data in the reference frame, that is:

$$\min_{\mathbf{R},\mathbf{S},\mathbf{t},p} \left(\sum_{i_1}^{N_k} \left\| \mathbf{R}\mathbf{S}\mathbf{x}_i^{(k)} + \mathbf{t} - \mathbf{x}_{p(i)}^{(1)} \right\| \right), \tag{2}$$

being $\mathbf{R}$ an orthogonal matrix, $\mathbf{S}$ a diagonal scale matrix, $\mathbf{t}$ a translation vector and $p : N_k \mapsto N_1$ a function that maps indexes of facial points across the the k-th and 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 [16]. Fig. 2 shows facial data acquired in two sample frames before and after the application of the adopted ICP procedure.

2.3 Face Sampler

Once facial data from the different frames are aligned to the data in the first frame—used as template—then resampling by interpolation is operated. The goal of this module is to compute a high-res image on the uniformly spaced grid Φ. However, as a result of the alignment of the generic k-th frame to the first one under the effect of Eq. 2, samples on the originally uniform grid Ω distribute irregularly. Therefore, it is necessary to convert this non-uniform raster to a uniformly spaced grid, and this is performed by way of a scattered data interpolation model based on Delaunay triangulation [17, 18]. The interpolation model acts as a function Γ that given the set of N_k scattered points $\left\{ \mathbf{R}^k \mathbf{S}^k \mathbf{x}_i^{(k)} \mathbf{t}^k \right\}_{i=1}^{N_k}$ that are expected to sample a 2D surface in the 3D space, projects this dataset onto a reference plane Π (the (x, y) plane of the first frame) and then estimates the *height* value of the surface for a generic point $p \in \Pi$ within the convex hull of the projected dataset (see Fig. 3).

In this way, given the super-resolution uniformly spaced grid Φ in Π, it is possible to estimate the value of the 2D surface for each point of Φ enclosed within the convex hull of the projection of the scattered points onto Π. This procedure is operated for each one of the N acquired depth frames so that for each point of Φ, N observations are available. The median of these observations is the estimated value of the superface on the super-resolution grid Φ.

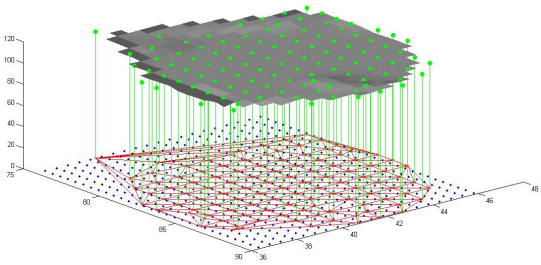


Fig. 3. Projection of data points of a generic frame onto the reference plane associated to the first frame distribute irregularly. Estimation of values of the underlying surface (shown in gray) on a regular grid (blue points) is accomplished by defining a Delaunay triangulation (red lines) on the projection of data points (red stars) and then interpolating the value of the surface within each triangle

3 Experimental Results

To the best of our knowledge, public data sets that provide, at the same time, sequences of low resolution face scans acquired with 3D consumer cameras, and high resolution 3D scans of the same subjects are not available. So, to bypass the lack of benchmark data and test our super-resolution approach, we constructed a proprietary data set which is released to the research community for comparative evaluations (see Sect. 3.1 for details). This data set in used in the tests reported in Sect. 3.2.

3.1 Data Set

In order to experiment the applicability and accuracy of our 3D super-resolution approach, we collected a test data set comprising low-resolution and high-resolution 3D scans. Currently, the data of 20 subjects are included while the subjects enrolling is still going on (we aim a number of about 50 subjects be comprised in the data set). In particular, for each person we captured:

- A 3D high-resolution face model acquired with the *3dMD* scanner. The model comprises a 3D 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 better, depending on the exact pre-calibration and configuration. All 3D models are provided in VRML format;
- A depth-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 also asked to slightly move the face around the yaw axis up to an angle

of about 60-70 degrees, so that both the left and right side of the face are visible to the sensor. This results in video sequences lasting approximately 10 to 15 seconds on average. Each depth video is released as a sequence of frames in PNG format and 16 bits gray scale.

The data are released in the same form they are acquired by the sensors, without any processing or annotation[1].

3.2 Error Measures

In order to evaluate the accuracy of the super-resolution process, we compared the 3D geometry of reconstructed face models of sample subjects against the corresponding 3D face models of the same subjects acquired with a high-resolution 3D scanner. This is similar to the problem of measuring the geometric distance between high and low resolution versions of a same triangular mesh, which is a common task in 3D mesh processing providing an indication of the quality of the simplification process that reduces the number of triangles [19]. In our work, the reconstructed mesh of a face originated by the super-resolution process can be regarded as a less accurate 3D representation of a same face acquired with the high-resolution scanner. According to this, we propose to use the error measure introduced in [20], based on the computation of the *Hausdorff* distance.

Given two surfaces S and S', the distance between a point p on S and the surface S' is defined as:

$$d(p, S') = \min_{p' \in S'} d(p, p'), \tag{3}$$

where $d(p, p')$ is the Euclidian distance between two points in S and S', respectively. The geometric distance, also called one-sided or single-sided *Hausdorff* distance, between two surfaces S and S' is then defined as:

$$d(S, S') = \max_{p \in S} d(p, S'). \tag{4}$$

This distance is not symmetric (i.e., $d(S, S') \neq d(S', S)$), so that it can underestimate the real distance between two surfaces. Due to this, a more accurate measure of the distance is obtained by using the symmetrical *Hausdorff* distance:

$$d_H(S, S') = \max\{d(S, S'), d(S', S)\}. \tag{5}$$

The point-to-surface distance of Eq. (3) is also used to define the *mean distance* d_m between two surfaces as the distances between points on S and the surface S', divided by the area of S:

$$d_m(S, S') = \frac{1}{|S|} \sum_{p \in S} d(p, S'). \tag{6}$$

[1] The data set can be accessed at the following link: `http://www.micc.unifi.it/datasets/4d-faces/`

The symmetric version of the mean distance is then defined as the average between the two single-sided mean distances, that is, $\bar{d}(S, S') = (d_m(S, S') + d_m(S', S))/2$. Practically, in computing Eq. (4) vertices of the mesh are used as sampling points p of the surface S. In addition, the *Root Mean Square* error (RMS) on the vertices of the two comparing meshes is also computed.

In our experiments, we used the data set described in the previous Section. As an example, Fig. 4 shows the 3D reference frame (i.e., the first frame of the depth-sequence) as triangulated mesh, the 3D reconstructed frame and the 3D high-resolution scan of three sample subjects (named, respectively, #1, #3, and #7) included in the data set. Before evaluating the distance between two face models, they are cropped (i.e., only the points included in a sphere centered on the nose tip and with 95mm of radius are retained), normalized with respect to their center of mass and aligned each other using the ICP algorithm.

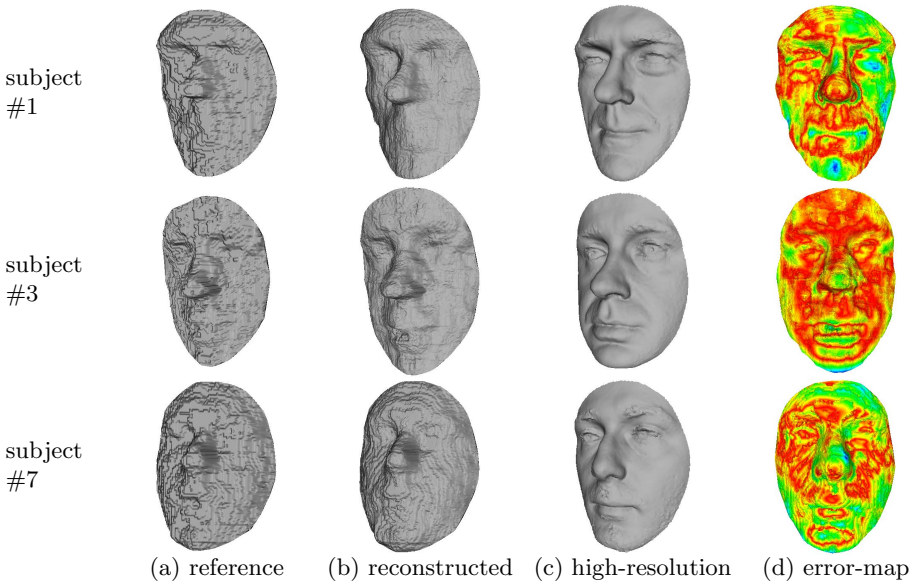


Fig. 4. Each row reports, for a different sample subject: (a) the 3D reference frame; (b) the 3D reconstructed frame; (c) and the 3D high-resolution scan. In (d), the error map of the Hausdorff distance is reported for the reconstructed model, where the error increases along the red-green-blue color scale.

Results are summarized in Tab. 1. In particular, we reported the average values for the symmetric *Hausdorff* distance (D_H), the symmetric *mean distance* ($\bar{d}$), and the *RMS* error computed between the high-resolution scan and, respectively, the reconstructed frame and the reference frame. In this way, a quantitative evidence of the increased quality of the reconstructed frame with respect to the reference one is obtained. The percentage variation of these error measures when passing from the reference to the reconstructed frame are also reported.

As general behavior, it can be observed that all the three error measures decreases when evaluated on the reconstructed frame instead of the reference one, with a percentage reduction of the error which varies from around 16% up to 23%, respectively, for the symmetric Hausdorff (d_H) and the RMS error. Fig. 4(d) also reports the error map of the Hausdorff distance, where the red-green-blue colors are associated to errors of increasing magnitude. In general, it can be seen that the error is small, with just a few areas of the reconstructed face colored in blue.

Table 1. The average distance measures computed between the 3D high-resolution face scan and, respectively, the reconstructed and the reference frame of each subject. The percentage variation of the distance between the errors for the reconstructed and reference frames is also reported.

average distance	d_H	$\bar{d}$	RMS
reference	14,508231	2.171087	3.112833
reconstructed	12.200201	1.682568	2.393260
% variation	-15.91%	-22.5%	-23,12%

4 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 3D frame using the scaled ICP algorithm, and estimating the statistics of the values in the low-resolution models in corresponding points. Preliminary qualitative and quantitative experiments have been performed on an acquired dataset that includes, for each subject, a sequence of low-resolution 3D frames and one high-resolution 3D scan used to provide 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 superface and the ground truth.

Acknowledgment. The authors acknowledge Lorenzo Seidenari for his valuable contribute to the design and development of the software modules for 3D data acquisition through the Kinect scanner.

References

1. Wang, Y., Liu, J., Tang, X.: Robust 3D face recognition by local shape difference boosting. IEEE Trans. on Pattern Analysis and Machine Intelligence 32, 1858–1870 (2010)

2. Berretti, S., Del Bimbo, A., Pala, P.: 3D face recognition using iso-geodesic stripes. IEEE Trans. on Pattern Analysis and Machine Intelligence 32, 2162–2177 (2010)
3. Colombo, A., Cusano, C., Schettini, R.: Gappy PCA classification for occlusion tolerant 3D face detection. Journal of Math. Imaging and Vision 35, 193–207 (2009)
4. Passalis, G., Perakis, P., Theoharis, T., Kakadiaris, I.A.: Using facial symmetry to handle pose variations in real-world 3D face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 1938–1951 (2011)
5. Huang, T., Tsai, R.: Multi-frame image restoration and registration. In: Advances in Computer Vision and Image Processing, vol. 1, pp. 317–339 (1984)
6. 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, 1621–1633 (1997)
7. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 1167–1183 (2002)
8. Farsiu, S., Robinson, M., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Trans. on Image Processing 13, 1327–1344 (2004)
9. Ebrahimi, M., Vrscay, E.: Multi-frame super-resolution with no explicit motion estimation. In: Proc. of Int. Conf. on Image Processing, Computer Vision, and Pattern Recognition (IPCV), pp. 455–459 (2008)
10. Smelyanskiy, V.N., Cheeseman, P., Maluf, D.A., Morris, R.D.: Bayesian super-resolved surface reconstruction from images. In: Proc. of IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), pp. 375–382 (2000)
11. Peng, S., Pan, G., Wu, Z.: Learning-based super-resolution of 3D face model. In: Proc. of IEEE Int. Conf. on Image Processing (ICIP), vol. II, pp. 382–385 (2005)
12. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: Lidarboost: Depth superresolution for tof 3D shape scanning. In: Proc. of IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), pp. 343–350 (2009)
13. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: Proc. of IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
14. Pan, G., Han, S., Wu, Z., Wang, Y.: Super-Resolution of 3D Face. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 389–401. Springer, Heidelberg (2006)
15. Arun, K., Huang, T., Blostein, S.: Least-squares fitting of two 3-D point sets. IEEE Trans. on Pattern Analysis and Machine Intelligence 9, 698–700 (1987)
16. 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, 442–452 (2010)
17. Faugeras, O.: Three-dimensional computer vision: A geometric viewpoint. MIT Press, Cambridge (1993)
18. Powell, M.: A review of methods for multivariable interpolation at scattered data points. Cambridge University Press (1996)
19. Aspert, N., Santa-Cruz, D., Ebrahimi, T.: MESH: Measuring errors between surfaces using the Hausdorff distance. In: Proc. of the IEEE Int. Conf. on Multimedia and Expo, vol. I, pp. 705–708 (2002)
20. Cignoni, P., Montani, C., Scopigno, R.: A comparison of mesh simplification algorithms. Computers & Graphics 22, 37–54 (1998)