Vis Comput (2014) 30:1275–1292
DOI 10.1007/s00371-014-0932-7

ORIGINAL ARTICLE

Selecting stable keypoints and local descriptors for person identification using 3D face scans

Stefano Berretti · Naoufel Werghi · Alberto del Bimbo · Pietro Pala

Published online: 8 April 2014

Abstract 3D face identification based on the detection and comparison of keypoints of the face is a promising solution to extend face recognition approaches to the case of 3D scans with occlusions and missing parts. In fact, approaches that perform sparse keypoints matching can naturally allow for partial face comparison. However, such methods typically use a large number of keypoints, locally described by high-dimensional feature vectors: This, combined with the combinatorial number of keypoint comparisons required to match two face scans, results in a high computational cost that does not scale well with large datasets. Motivated by these considerations, in this paper, we present a 3D face recognition approach based on the meshDOG keypoints detector and local GH descriptor, and propose original solutions to improve keypoints stability and select the most effective features from the local descriptors. Experiments have been performed to assess the validity of the proposed optimizations for stable keypoints detection and feature selection. Recognition accuracy has been evaluated on the *Bosphorus* database, showing competitive results with respect to existing 3D face identification solutions based on 3D keypoints.

Keywords 3D face recognition · 3D Keypoints detection · Stable scale space selection · Feature selection

S. Berretti (✉) · A. del Bimbo · P. Pala
Department of Information Engineering,
University of Florence, Florence, Italy
e-mail: stefano.berretti@unifi.it

N. Werghi
Department of Electrical and Computer Engineering,
Khalifa University of Science, Technology and Research,
Sharjah Campus, Abu Dhabi, United Arab Emirates
e-mail: naoufel.werghi@kustar.ac.ae

1 Introduction

Recognition of persons' identity using 3D scans of the face has been proposed as an alternative or complementary solution to conventional face recognition approaches that exploit the appearance of the face in 2D still images or videos. In fact, since 3D face scans capture the full 3D geometry of the face, they are expected to feature less sensitivity to lighting conditions and pose variations, thus allowing accurate face recognition also in real-world applications. Though consolidated 3D face recognition solutions exist that achieve high accuracy in cooperative scenarios (see the survey in [7], and the literature review in [3,34]), solutions enabling face recognition with uncooperative settings are now attracting an increasing interest [26]. In semi-cooperative or uncooperative scenarios, probe scans are acquired under unconstrained conditions that may result in face scans with *missing parts* (due to non-frontal pose of the face) or *occlusions* (due to hair, glasses, scarves, hand gestures covering the face, etc.), thus demanding for methods capable of performing recognition just using parts of a 3D face scan. Indeed, works focusing on this topic are still preliminary also due to the limited number of 3D face databases that include partial/occluded acquisitions of the face. From this point of view, the release of the *Bosphorus* dataset [31] has boosted the research in this direction, providing a large reference benchmark to evaluate 3D face recognition methods in the case of face scans with expressions, missing parts and occlusions.

Some recent works attacked the problem of 3D face recognition in the presence of occluded/partial scans, starting from the observation that describing the face with local geometric information extracted in the neighborhood of 3D keypoints can naturally allow partial face comparison by sparse keypoints matching. In particular, such a framework has been experimented either considering 2D solutions, which exploit

SIFT-like detectors applied to 2D-maps of the face (e.g., depth maps or 2D-maps of some face descriptor) [5,16], or directly detecting keypoints of 3D meshes of the face [21,32]. Though effective results have been obtained by the methods which apply detectors to 2D maps of the face, these solutions are limited by the need of a preliminary accurate registration of the face scans to a common reference system, which is by itself a difficult task requiring in many cases the detection of facial landmarks. So, in perspective, methods based on 3D keypoint detectors appear more promising, in that they do not require any preliminary alignment of the mesh and can exploit its full 3D geometry.

In general, following the definition given in [33], 3D keypoints (or interest points) are prominent points of a shape according to a particular definition of interestingness or saliency: They are extracted by a 3D detector, which analyses local neighborhoods around the elements of a given surface. In the case of 3D faces, it is relevant to evidence the difference between *keypoints* and *landmarks*. Facial landmarks are points of the face defined according to anatomical studies of the facial bones and muscles, and correspond to visible or palpable features (skin or bones) of the face [9]. Their localization represents a precursor operation for several 3D face analysis applications, such as face recognition, facial expression recognition, face registration and reconstruction, etc. For example, in his anthropometric studies, Farkas defined as many as 47 main facial landmarks, and used them for facial measurements [14]. Due to their characteristics, the localization of facial landmarks in 3D scans can exploit the knowledge of the human face, so that they can be manually annotated by a human operator. However, few of them can be detected automatically in a reliable way. For example, in the state of the art approach by Perakis et al. [28], a method is proposed to automatically detect eight landmarks (eye and mouth corners, nose and chin tips) on 3D facial scans that exhibit yaw and expression variations. The main contribution of the method is its applicability to large yaw variations (up to 82°) that often result in missing (self-occluded) facial data, and its tolerance against varying facial expressions. Candidate landmarks are first detected by exploiting the 3D geometry-based information of *shape index* and *spin images*. The candidate landmarks are then identified and labeled by matching them with a Facial Landmark Model of facial anatomical landmarks. The method is extensively evaluated against a variety of 3D facial databases and achieves state-of-the-art accuracy (4.5–6.3 mm mean landmark localization error).

Differently, 3D keypoints of the face are detected based only on the geometric properties of the surface, without any additional knowledge about the context, and without requiring them to be located in correspondence to any specific anatomical trait of the face. As a consequence, many keypoints are typically detected, and they are located in scattered positions across the face, so that they cannot be manually annotated by a human operator. According to this, the most relevant aspect of a keypoints detector is the possibility to extract keypoints in repeatable positions across different scans of a same face under a number of nuisances that can affect the input data. Based on these considerations, coincidence between landmarks and keypoints of 3D face scans is possible just for the few landmarks characterized by highly distinctive geometric variations that make them detectable also as keypoints (e.g., this can be the case for the eyes and mouth corners). In general, there is not any evident relationship between landmarks and keypoints of 3D face scans and they are typically detected following different approaches and are used for different purposes.

In the following, methods that rely on 3D keypoints detection and description for face recognition are summarized.

1.1 Related work

To the best of our knowledge, a few works used 3D keypoints detectors for face recognition, using different descriptors and matching policies.

The solution proposed in [24] was the first to exploit 3D keypoints of the face, by considering as extrema the vertices of the face mesh for which the difference between the maximum and minimum eigenvalues of a local principal components analysis is over a given threshold. However, this work was tested on the FRGC v2.0 dataset that does not include occluded/partial scans. In [23,32], the framework of SIFT keypoints detector has been reformulated to operate on 3D face meshes by defining the meshSIFT detector and local descriptor. This first performs a scale-space analysis of the mesh through subsequent smoothing of the 3D geometry, and then identifies as 3D keypoints the local extrema of the mean curvature extracted from the smoothed versions of the original mesh through the scale. Local descriptors are also defined at the keypoints using nine local regions (arranged according to a daisy-like pattern) and computing for each of them a pair of histograms (the shape index and the angle between surface normals are used). The meshSIFT has been used as keypoints detector also in [21]. In this case, a quasi-daisy local shape descriptor of each feature point has been obtained using multiple order histograms of differential quantities extracted from the surface (these include gradient, shape index and gradient of shape index). During the match, these local descriptors are compared using the angle among them. The approach in [6] used the meshDOG as detector of relevant extrema of the face surface. After keypoints detection, different local descriptors have been extracted at the keypoints and their performance has been evaluated and compared. All the approaches in [21], [23,32] and [6] have been experimented on the *Bosphorus* dataset, showing high face recognition accuracies.

A common limitation of the solutions above is represented by the large number of detected keypoints. In general, this is reported to be a positive and required behavior of keypoints-based solutions, in that it is expected to increase the number of valid keypoint correspondences across matching scans [33]. However, the larger the number of keypoints, the greater is the probability to detect also unstable keypoints. This has a twofold effect: on the one hand, the presence of unstable keypoints is likely to increase the number of noisy keypoints matching; on the other hand, the number of keypoints matching grows quadratically with the keypoints, thus resulting in a demanding computational cost (this latter effect being further exacerbated by the high dimensionality of local keypoint descriptors).

In fact, none of these keypoint detectors addressed the problem of how to select the 3D keypoints that are most significant, relating at the same time the significance of the keypoints with the stability of the corresponding descriptors. Moreover, the different relevance of individual features of local descriptors is not considered, and solutions to reduce the dimensionality of the descriptors are not considered (for example by selecting just the most relevant features). Actually, a combined analysis of these aspects would have the potential of finding a better compromise between the number of keypoints, the features included in the local descriptors and the accuracy of recognition.

1.2 Contribution and paper organization

Moving from the considerations above, in this work we propose and experiment original solutions which aim to address the current limitations of 3D face recognition methods based on 3D keypoints. To this end, we first report about a 3D face recognition method that we recently proposed in [6], which is capable of performing subjects identification also in the presence of facial expressions, occlusions and missing parts of the face. In particular, the approach extracts meshDOG keypoints and locally describes the face around them using a *multi-ring* Geometric Histogram. Robust keypoint correspondences are then obtained in the match of two faces using outliers rejection with the RANSAC algorithm. We elaborate on such approach in several ways. On the one hand, we provide an improved keypoints extraction procedure that permits the selection of distinctive keypoints, with the advantage of a lower computational cost. On the other, we propose an original analysis, which permits: (i) improving the keypoints detection by relating the scale of the keypoints to the stability of the local descriptor; (ii) reducing the number of keypoints by accounting for their distribution and clustering; (iii) identifying the most relevant features of the local keypoint descriptor through a feature selection analysis. Though presented for the meshDOG keypoints detector and GH descriptor, these solutions can be regarded as general methods that, in principle, can be applied to different 3D keypoints detectors and descriptors as well, thus covering a broad range of applicability. In so doing, we emphasize that the main goal of this work is to address 3D face recognition in the case of static high-resolution scans with expressions, missing parts and occlusions. Our solution is not targeted for the new generation of dynamic low-resolution low-cost consumer cameras, like Kinect. These devices are likely to produce 3D dynamic sequences with missing parts and occlusions, but at a resolution and noise level which are currently not addressable by our solution. Ad-hoc methods should be used in this case, like those proposed in recent works [15,20,25].

The rest of the paper is organized as follows: The face representation based on meshDOG detector and *multi-ring* geometric histogram (GH) descriptor is summarized in Sect. 2, together with an effective face comparison algorithm that performs outliers removal using RANSAC; Investigation of the keypoints stability and relevance, and the selection of the most relevant features of the local descriptors are reported in Sect. 3. Evaluation and comparison of our work with respect to state of the art solutions on the *Bosphorus* dataset are given in Sect. 4; In the same section, we report verification results on the FRGC dataset, and provide evidence on the effect that noisy scans have on recognition. Finally, results and future research directions are discussed in Sect. 5.

2 Face representation and comparison

In the following, we present a framework for representing and comparing 3D scans of the face, which is based on meshDOG keypoints detection and local description using GHs. This framework was originally proposed in our previous work [6]: Here, it is modified and extended. In particular, the keypoints detection is modified through an improved scale-space approach for keypoints detection, which allows reducing the number of scale levels (and so their computation time) without affecting the number of keypoints and their repeatability. Further, the approach is extended and substantially improved through an original investigation that permits the selection of stable keypoints and of the most discriminating features of the local descriptor, as detailed in Sect. 3. In so doing, a subset of 80 subjects of *Binghamton University* 3D facial expression database (BU-3DFE) [36] has been used as training dataset. For each subject, six different facial expressions at four gradations (from moderated to exaggerated), plus the neutral one have been considered (25 scans per subject, 2,000 scans in total). This permitted us to keep separated the dataset used in the development from those used in the face recognition experiments (i.e., the *Bosphorus* and the FRGC datasets).

2.1 3D keypoints detection

Among 3D keypoint detectors [8,33], the meshDOG algorithm [37] has proven its effectiveness in locating repeatable extrema on 3D meshes [8]. In the following, the method is adapted to the particular case of extracting keypoints from 3D meshes of the face.

Given a 3D mesh $\mathcal{M}$, a scalar function $f(v) : \mathcal{M} \rightarrow \mathcal{R}$ can be defined, which returns a scalar value for any vertex $v \in \mathcal{M}$. In our case, we defined $f(v)$ as the *mean* curvature at vertex v, computed according to [29]. Though such function is not completely invariant to local isometric deformations, the keypoints detected using this function turned out to be more stable on 3D face data than keypoints obtained using Gaussian curvature. Similar results were also reported in [33] for meshDOG and in [23,32] for meshSIFT. Once the function f is computed for each vertex of the mesh, the keypoints detection proceeds in three steps.

Scale-space In the first step, a *scale-space* representation of the function f is constructed. At each scale t, the function f at vertex v_i is convolved with a Gaussian kernel, which depends on the scale through $\sigma(t)$:

$$g_{\sigma(t)}(x) = \frac{1}{\sigma(t)\sqrt{2\pi}} \cdot \exp(-x^2/2\sigma(t)^2), \tag{1}$$

being $x = \|\mathbf{v}_j - \mathbf{v}_i\|$, with vertices v_j confined to the neighborhood rings of vertex v_i. In particular, the ring of a vertex $rg(v, n)$ is the set of vertices that are at distance n from v on $\mathcal{M}$, where the distance n is the minimum number of edges between two vertices. Thus, $rg(v, 0)$ is v itself and $rg(v, 1)$ is the set of direct neighbours of v. The neighbourhood $Nn(v)$ is the set of rings $\{rg(v, i)\}_{i=0,\ldots,n}$. In the practice, $t = \{1, 2, \ldots, o \cdot s\}$, using $o = 4$ octaves and covering each octave in $s = 6$ steps (24 scales in total), so that $\sigma(t) = 2^{\frac{1}{s-2} \cdot \lceil t/s \rceil} \cdot e_{\text{avg}}$, with e_{avg} the average length of mesh edges.

Using the convolution kernels defined above, the scale-space of f is built incrementally on $L+1$ levels, so that: $f_0 = f$, $f_1 = f_0 * g_{\sigma(1)}$, $f_2 = f_1 * g_{\sigma(2)}, \ldots, f_L = f_{L-1} * g_{\sigma(L)}$. The difference of Gaussian (DOG) is then obtained from the difference of adjacent scales, e.g., $\text{DOG}_1 = f_1 - f_0$, $\text{DOG}_2 = f_2 - f_1, \ldots, \text{DOG}_L = f_L - f_{L-1}$ ($L = o \cdot s$, DOG is computed in total). In so doing, it is relevant to note that the geometry of the face does not change, but the different scalar functions f_k and DOG_k defined on the mesh. Once the scale space is computed, the initial set of extrema is selected as the maxima of the DOG across scales.

An example of the scale-space construction is reported in Fig. 1. In (a), f_k values at different scales (f_0 being the mean curvature) are shown for a sample face. In (b), gray levels are used to represent the DOG values at different scales (scales 2, 8, and 16 are reported).

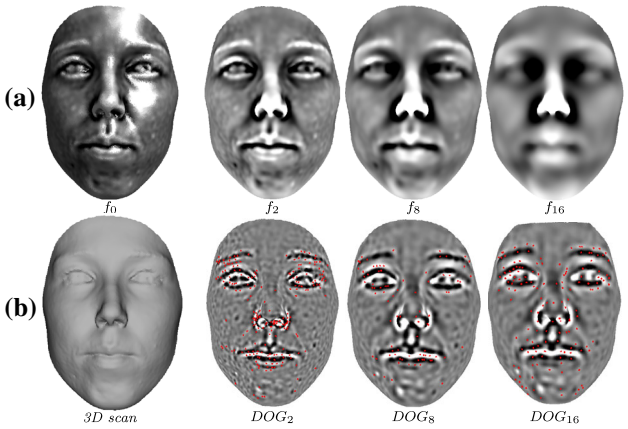


Fig. 1 **a** Values of f_k at different scales (f_0 is the mean curvature); **b** 3D face scan and DOG_k values at different scales (detected keypoints at each scale are depicted in *red*)

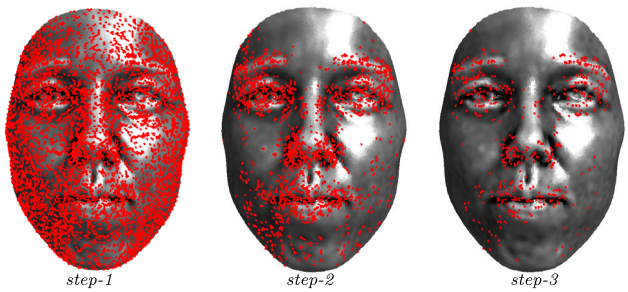


Fig. 2 Extrema points detected after each of the three steps of the keypoints detection algorithm

Percentage threshold In the second step, the extrema of the scale space are sorted according to their magnitude, and only the top 4 % are retained.

Corner analysis In the last step, unstable extrema are removed by retaining only those with corner characteristic, according to the Hessian computed at each vertex v of the mesh [22]. The ratio between the maximum λ_{max} and the minimum λ_{min} eigenvalues of the Hessian matrix is a good indication of a corner response, which is independent of the local coordinate frame ($\lambda_{\text{max}}/\lambda_{\text{min}} = 10$ has been used as a minimum value of threshold responses).

Figure 2 shows the extrema detected at the end of each of the three steps of the detection algorithm.

2.2 Multi-ring geometric histogram

Local description of the surface at the keypoints is obtained by relying on the concept of ordered ring facets (ORFs) and using geometric histogram (GH) [2].

Given a central facet t_c at which a keypoint has been detected, the ORF defines a ring-wise neighbourhood through a sequence of concentric rings of facets emanating from t_c. The facets are arranged circular wise within each ring and the

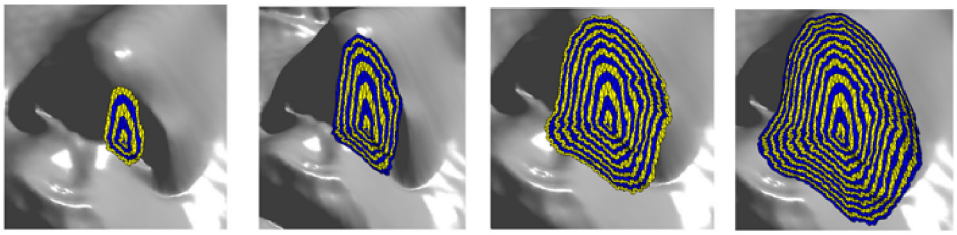


Fig. 3 ORFs with different neighbourhood size constructed at a facial keypoint

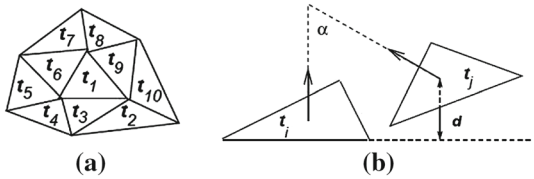


Fig. 4 GH computation: **a** Central facet t_1 and its neighbour facets; **b** Geometric measurements used to characterize the relationship between two facets t_i and t_j. For each pair (t_1, t_j) in (**a**), the angle α between the two facets' normals, the minimal and the maximal of the perpendicular distance from the plane of t_1 to the facet t_j are computed. The pairs (α, d) derived from these measurements are accumulated in a 2D matrix so as to obtain a geometric distribution

size of the neighbourhood is simply controlled by the number of rings. When the triangular mesh is regular and the facets are nearly equilateral, the ORFs approximate iso-geodesic rings around the central facet t_c. The ORFs are constructed with linear complexity (detailed algorithms for ORF computation are given in [35]). As an example, Fig. 3 shows ORFs with increasing neighbourhood size.

Let us consider a triangular mesh approximation $\hat{M} = \{t_1, \ldots, t_M\}$ of a 3D surface. Given a central triangular facet t_i, the basic idea of the GH is to construct a discrete geometric distribution, which describes the pairwise relationship between t_i and each of the surrounding facets within a predefined neighbourhood. The range of the neighbourhood controls the degree to which the representation is a local description of shape.

Figure 4 shows the measurements used to characterize the relationship between a central facet t_i and one of its neighbouring facets t_j. These measurements are the relative angle α between the facet normals; the range of perpendicular distances d from the plane in which the facet t_i lies to all the points on the facet t_j. The range of perpendicular distances is defined by $[d_{\min}, d_{\max}]$, where $d_{\min}$ and $d_{\max}$ are, respectively, the minimal and the maximal of the distance from the plane in which t_i lies to the facet t_j. In practice, these values are obtained by calculating the distances to the three vertices of the facet t_j and then selecting the minimal and the maximal distance. Since the distance measurement is a range, a single value $d_{\min} \leq d \leq d_{\max}$ is derived, based on the amplitude of the range $[d_{\min}, d_{\max}]$ and the resolution used for distance quantization. The set of pairs (α, d) computed between a given facet and its neighbours are entered to a 2D discrete frequency accumulator that encodes the perpendicular distance d and the angle α. This accumulator has size $I \times J$, where I and J are the number of quantization (bins) for α and d, respectively (8 and 10 in our case). Finally, values of the accumulated matrix are normalized so as to sum up to 1. The resulting distribution is invariant to rigid transformations of the surface and is also stable in the presence of surface clutter and missing data. In our approach, a GH is constructed in an incremental way for each of the rings in the ORF of a keypoint. According to this, a multi-ring GH (mr-GH) is obtained as a set of GHs constructed on the sequence of rings which surround a keypoint. This improves the descriptiveness of the GH by capturing information on how the local surface changes at increasing distance from the keypoint.

In the comparison of any two mr-GHs, the normalized GH is regarded as a probability density function, and the Bhattacharyya distance (d_B) is used as metric for evaluating the similarity between GHs at each ring. According to this, given two GHs in the form of linear arrays with $K = I \times J$ elements, $A(l) = \{a_1, \ldots, a_K\}$ and $B(l) = \{b_1, \ldots, b_K\}$, their distance at ring-l is computed as:

$$d_B(A(l), B(l)) = \sqrt{1 - \sum_{k=1}^{K} \sqrt{(a_k \cdot b_k)}}. \quad (2)$$

The overall distance between two mr-GHs computed on L rings is then obtained by accumulating the distances between the GHs at different rings.

2.3 Face comparison

Given two face scans, their comparison is performed by matching the mr-GHs of keypoints under the constraint that a consistent spatial transformation exists between inlier pairs of matching keypoints.

To this end, the mr-GHs at the keypoints detected in probe and gallery scans are compared using the accumulated Bhattacharyya distance so that, for each keypoint in the probe, a candidate corresponding keypoint in the gallery is identified. In particular, a keypoint k_p in the probe is assigned to a keypoint k_g in the gallery, if they match each other among all keypoints, that is if and only if k_p is closer to k_g than to any other keypoint in the gallery, and k_g is closer

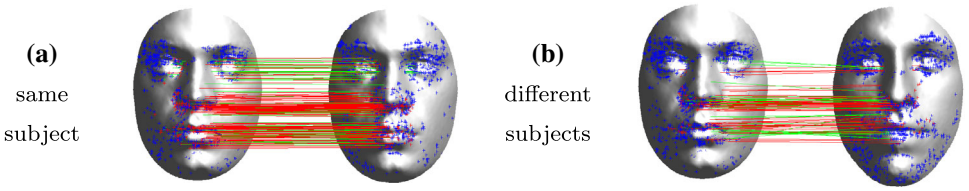


Fig. 5 Comparison between face scans: **a** same subject; **b** different subjects. The detected keypoints are shown with a *plus symbol* ("+") colored in *blue*. *Green lines* indicate matching keypoints, while *red lines* are the inliers matching after RANSAC

to k_p than to any other keypoint in the probe. In so doing, it is also required that the second best match is significantly worse than the best one, i.e., a match is accepted if the ratio between the best and the second best matches is lower than 0.7. The use of this threshold follows the idea appeared in other methods based on finding correspondences between pairs of keypoints (i.e., it was first proposed by Lowe in the SIFT matching algorithm for 2D still images [22]). As a result of the match, a candidate set of keypoint correspondences is identified: The actual set of correspondences is then obtained using the RANSAC algorithm [38] to remove outliers. This involves generating transformation hypotheses using a minimal number of correspondences and then evaluating each hypothesis based on the number of inliers among all features under that hypothesis. In our case, we modeled the problem of establishing correspondences between sets of keypoints detected on two matching scans as that of identifying points in $\Re^3$ that are related via a *rotation*, *scaling* and *translation* (RST) transformation. According to this, at each iteration, the RANSAC algorithm validates sampled pairs of matching keypoints under the current RST hypothesis, updating at the same time the RST transformation according to the sampled points. In this way, corresponding keypoints whose RST transformation is different from the final RST hypothesis are regarded as outliers and removed from the match. The number of keypoint matches is then used as similarity measure between two scans.

Matching examples are reported in Fig. 5a, b, for scans of same and different subjects, respectively. In this figure, detected keypoints are marked with "+" and highlighted in blue; corresponding keypoints based on GH matching are connected by green lines; the inlier keypoint correspondences, which pass the RANSAC test, are shown with red lines. It can be observed as applying RANSAC, just the matches that show a coherent RST transformation among each other are retained, thus avoiding matches of keypoints that are located in different parts of the face in two scans.

The effect on recognition due to varying the threshold on the ratio between the best and the second best matches in a keypoints correspondence has been evaluated in a face identification experiment. In this experiment, the neutral scans of 80 subjects of the BU-3DFE database are included in the gallery, and the 12 scans at the first two gradations of expression intensity per subject are used as probes (960 probes in total). Results are reported in Fig. 6, where the rank-1 accuracy is plotted against the variation of the threshold (values from 0.4 to 1 are used, with step 0.1). Best results are achieved for the threshold equal to 0.7.

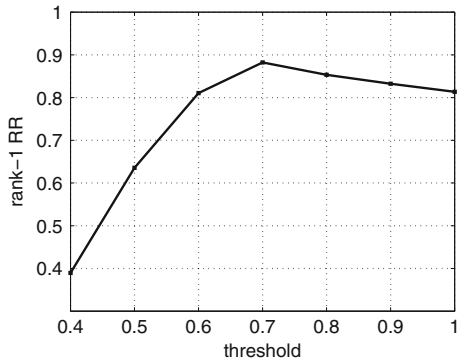


Fig. 6 BU-3DFE: Face identification accuracy as a function of the *threshold* on the ratio between the best and the second best matches for accepting a keypoint correspondence

3 Stable keypoints and descriptors

Previous works that use keypoints for 3D face recognition directly applied extracted extrema and their descriptors to perform face comparison, without any further analysis aiming to select stable keypoints and descriptors [5,24,32]. In practice, the effectiveness of the face representation is shown in these works by recurring to the overall accuracy of recognition. Actually, defining appropriate methods to identify stable keypoints and select the most effective features from the local descriptors could permit the optimization of the process of selection/description of the keypoints. In particular, three aspects are worth of investigation:

- Optimal scale selection for keypoints description;
- Keypoints distribution and clustering;
- Feature selection of local descriptors.

In the remaining part of this section, we present original methods addressing the three aspects above.

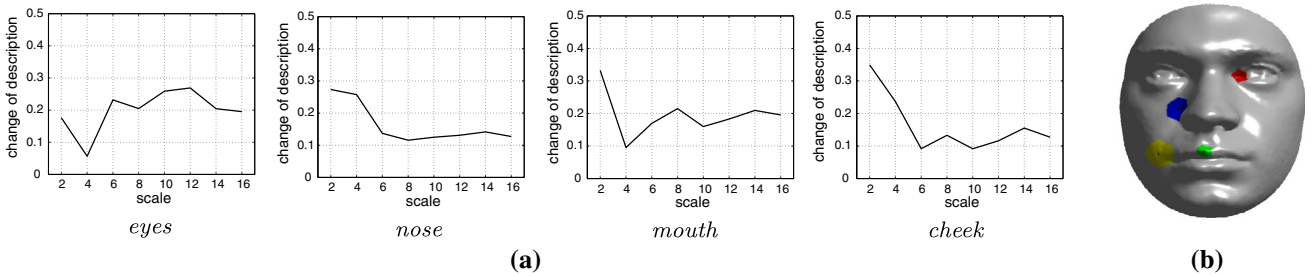


Fig. 7 Examples of scale selection: **a** The *graphs* show the change in the local description as a function of the scale of four keypoints; **b** The scale for which the functions in (**a**) have global minima is shown on the scan with *different colors* (*red*, *blue*, *green* and *yellow* are used, respectively, for keypoints in the *eyes*, *nose*, *mouth* and *cheek* region of the face)

3.1 Optimal scale selection for keypoints description

In the following, we propose an approach to optimize keypoints detection according to the stability of the local descriptor. The basic idea we follow here is to select the scale at which keypoints are detected, and consequently the extent of the local support for computing the GH descriptor, based on a concept of maximal stability of the descriptor throughout the scales. In practice, maximal stability is obtained when the difference between descriptors extracted for consecutive scales reaches a minimum. In this respect, our analysis shares some common concept with the approach proposed in [10] for SIFT on 2D still images. Therefore, the resulting detector uses the descriptor to select the characteristic scale. This is obtained in two steps: in the first step, the extrema are detected at multiple scales to determine informative and repeatable locations as reported in Sect. 2.1; in the second step, the characteristic scale for each location is selected by identifying maximally stable mr-GH descriptors. In so doing, we use description stability as criterion for scale selection: The scale for each location is chosen such that the corresponding mr-GH representation changes the least with respect to the scale.

Figure 7 illustrates some results of the proposed scale selection criterion for meshDOG keypoints and GH descriptor. Plots in (a) show how the descriptor changes as the scale (i.e., the number of rings) increases, for four keypoints located, respectively, in the *eyes*, *nose*, *mouth* and *cheek* region of a face scan. To measure the difference between mr-GHs at different scales, we used the accumulated Bhattacharyya distance as defined in Sect. 2.2. The absolute minimum of the function in each plot determines the scale at which the descriptor is the most stable for each keypoint. The scales (regions) selected for the four keypoints in (a) are depicted by color patches on the face scan in (b). From the figure, it can be observed as the selected scale changes for different keypoints. For example, in the eyes region it results that the best scale for computing the local descriptor is quite small (just four rings) due to the effect of noise, while the extent of the scale increases for the keypoints located in the nose and cheek regions (in this latter case, the scale extends up to ten rings).

3.2 Keypoints clustering and distribution

Observing the distribution of keypoints across different facial scans, it results that they are not fully dispersed but, on the opposite, a considerable portion of them shares quite close locations. This spatial clustering aspect of the keypoints distribution is attractive as it has the potential of reducing the combinatorial number of keypoint matches that occur in face comparison. In our case, a cluster of keypoints S_{k_c} is defined as the group of keypoints that are within the spherical neighborhood of radius r of a keypoint k_c of the set, that is:

$$S_{k_c} = \{k_i : |k_i - k_c| < r,\ i = 1, \ldots, n_{k_c},\ n_{k_c} > 1\}. \quad (3)$$

According to this, a set with a single keypoint cannot be considered as a cluster, but a cardinality greater than one is required. The radius is set to $r = \rho \cdot e_{\text{avg}}$, where e_{avg} is the average length of mesh edges and ρ is an integer. The computation of the clusters is performed with a neighborhood grouping procedure. The hypothesis we consider here is that keypoints in a cluster should present similar or at least quite close descriptors, so that just the central keypoint k_c can be considered in the matching, rather than all the keypoints in the cluster. The degree of compliance of a cluster with this assumption can then be adopted as a validity criterion on whether or not to use that cluster in the face comparison. For instance, if a cluster obtained with a low r has quite disparate descriptors at its keypoints, it might result in conflicting matching. Following this intuition, we examined the behavior of the clusters, at increasing radii of the sphere, in terms of: (i) variation of their number; (ii) homogeneity with respect to the GH descriptor that is the extent to which the local descriptors computed at keypoints in a cluster keep close as r increases (ideally, the descriptors should be identical for all the keypoints within a cluster). These two aspects have been captured by computing the following quanti-

Table 1 The four groups of clusters as a function of the values of the *intra-cluster* distance μ

Cluster group	μ
group-1	$\mu \leq 0.2$
group-2	$0.2 < \mu \leq 0.5$
group-3	$0.5 < \mu \leq 0.8$
group-4	$0.8 < \mu$

ties (with ρ ranging from 1 to 5, and thus for increasing radius r):

- The *keypoints reduction* $\alpha = (N - (\eta + \chi))/N$, where N, η and χ are the number of keypoints, clusters and isolated keypoints, respectively. Since the numerator represents the number of keypoints that would be used in the matching if a cluster is substituted by its central keypoint, the ratio α can be regarded as the amount of keypoints reduction due to clustering;
- The *intra-cluster distance* μ, that is the mean of the pairwise distances between the GH descriptors of any two keypoints in a cluster. Depending on the value of μ, clusters are divided into four groups as reported in Table 1. The threshold 0.2 has been chosen upon a statistics in which we estimate the maximum distance between GH descriptors computed at the same locations for different scans of a same person. This subdivision reflects the homogeneity of the GH descriptors within a cluster: clusters in *group-1* are the most homogenous; At the other extreme, clusters of *group-4* exhibit the highest disparity and thus should be discarded from the match.

Results of a statistical analysis performed using the above measures are summarized in Fig. 8. In Fig. 8a, the mean and standard deviation of the keypoints reduction α are plotted against the radii of the spherical neighborhood (values of ρ from 1 to 5 have been used). We notice that for $\rho = 1$, that is for clusters practically confined within a facet and its adjacent neighbors, the percentage of keypoints reduction is around 30 % on average. This value increases up to more than 50 % for $\rho = 2$. This is encouraging if we assume that the local shape is not expected to change too much in a such reduced neighborhood.

Figure 8b reports the percentage of clusters belonging to the four groups listed in Table 1, with respect to the spherical neighborhood size ρ. We notice that the average number of clusters having an intra-cluster distance μ in *group-1* remains above 60 % up to the third spherical neighborhood. This is encouraging as it means that the homogeneity, and thus the trustworthiness of the clusters, for a considerable number of clusters is not compromised when the extent of the neighborhood increases. On the other hand, we notice that the number of non-homogenous groups remains less than 30 % up to the second neighborhood size, especially for *group-2*. Extremely non-homogenous clusters (*group-3* and *group-4*) have very low proportions, yet they are present across all the neighborhood sizes. These can be viewed as outliers or instable clusters (*group-4* in particular), for which the descriptors show large disparity. Therefore, including in face comparison clusters belonging to *group-3* and *group-4* can jeopardize the recognition accuracy. This statistical analysis provides insights into the clustering of keypoints, evidencing the potential of exploiting this characteristic for reducing the combinatorial number of keypoints matching. The ultimate goal of this analysis is to eliminate unreliable keypoints, thus increasing the number of matches between corresponding keypoints. Based on the previous considerations and results, the keypoints are ranked in the following way: (i) keypoints belonging to clusters of *group-1*; (ii) isolated keypoints; (iii) keypoints belonging to clusters of *group-2*.

In the analysis above, the position of the clusters across different zones of the face is not accounted. Actually, this position can affect the relevance of the clusters in face comparison. This aspect has been investigated through a statistical analysis of the relationship between the spatial distribution of the clusters across the facial surface and the intra-cluster distance μ. A preliminary observation of the spatial

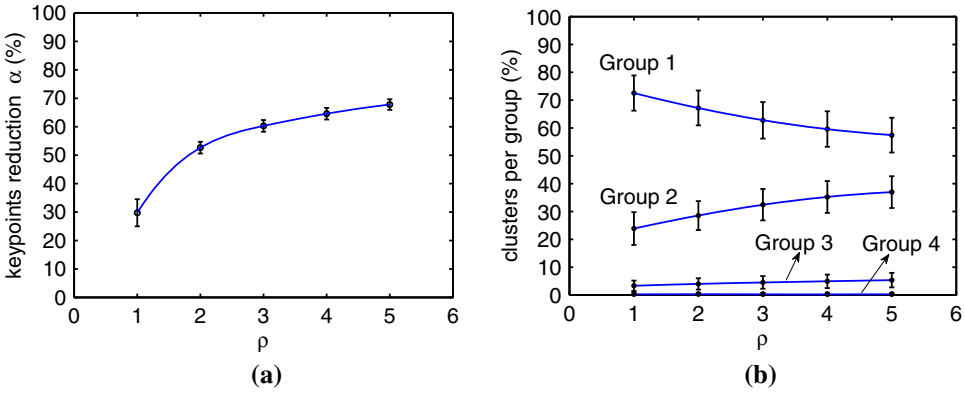


Fig. 8 **a** Mean keypoints reduction α, in percentage, with respect to the spherical neighborhood size (parameterized by ρ); **b** Percentage of clusters belonging to *groups* 1, 2, 3, and 4, for increasing spherical neighborhood size ρ. In both plots, the *vertical bars* report the standard deviation at each ρ

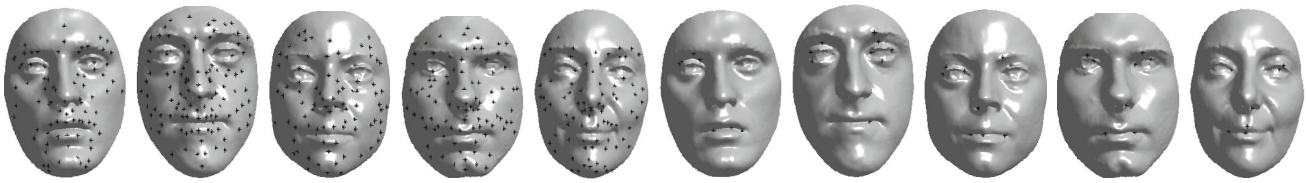


Fig. 9 Distribution of cluster centers for subjects in *group-1* (*top*) and *group-3* (*bottom*)

Table 2 Percentage of clusters in *group-3* and *group-4* that are distributed in different zones of the face for increasing values of ρ

	ρ				
	1	2	3	4	5
Lip (%)	23.4	23.1	27.3	25.0	22.9
Nose (%)	28.2	25.8	23.7	19.8	22.5
Eyes-eyebrows (%)	47.1	49.7	48.2	54.3	54.5
Other (%)	1.3	1.4	0.8	0.8	0.1

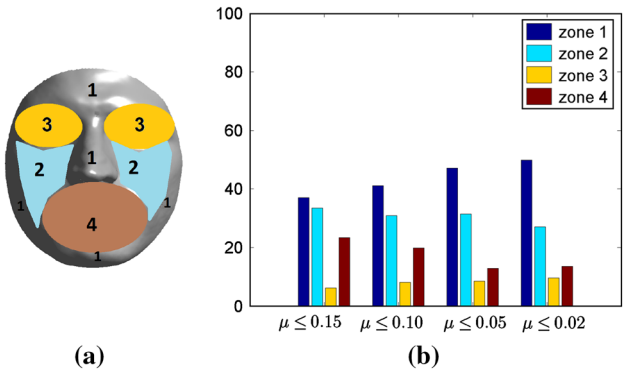


Fig. 10 **a** Facial zones numbered from 1 to 4 according to their increasing sensitivity to facial expressions; **b** The related clusters distribution (in percentage) across the different sub-groups of *group-1*

distribution of the four groups of clusters showed that clusters in *group-1* and *group-2* are spread over the entire face, whereas clusters in *group-3* and *group-4* are restricted to the *lips*, *eyes-eyebrows* and the *base of the nose*. Examples are given in Fig. 9 for *group-1* and *group-3*.

To investigate this aspect, we counted the number of clusters in *group-3* and *group-4* (merged together) located in each of the afore-mentioned zones of the face. The corresponding statistics is depicted in Table 2. The reported percentages confirm the visual observation, with about half of the clusters located in the eyes-eyebrows zone and the rest distributed in the lips and the base of the nose (more specifically, within and around the irregular mesh locations caused by nostrils).

A second statistical analysis was conducted on the clusters in *group-1* obtained with a spherical neighbourhood of $\rho = 3$. In this analysis, we considered four facial zones ranked according to their sensitivity to facial expressions, namely, the *nose* and the *border* of the face (*zone 1*, least affected by facial expressions), the *cheeks* (*zone 2*), the *eyes* and *eyebrows* (*zone 3*), and the *mouth* (*zone 4*, most affected by facial expressions). These zones are sketched in Fig. 10a. We computed the distribution of the clusters for four subranges of *group-1*, namely, $\mu \leq 0.15$, $\mu \leq 0.1$, $\mu \leq 0.05$, and $\mu \leq 0.02$. The corresponding statistics is depicted in Fig. 10b, showing the histograms of the number of clusters obtained in each zone and for the different sub-groups. The variation of the clusters distribution across the different subgroups shows an increase in the percentage of stable clusters in the *zone 1*, reaching more than 50 % for μ less than 0.05. An opposite behaviour is observed for the *mouth* (*zone 4*), where the related clusters get below 14 % for the same aforementioned range of μ. The number of clusters at the *cheek*, while significant, shows a relatively little variation. The same can be noticed for the *eyes* (*zone 3*), though it shows a little number of clusters across the different sub-groups.

Results suggested us that there is an association between the homogeneity of the clusters and the sensitivity to facial expressions (and so, to some zones of the face). This aspect is particulary noticeable at the extreme ranges of μ, where the most (respectively, least) stable clusters tend to be located in the zones of the face which are least (respectively, most) sensitive to facial expressions. This also suggested us that the quantity μ of a cluster reflects, to some extent, the likelihood of that cluster of being located at a particular zone of the face and, therefore, can be utilized for establishing plausible correspondences in the face matching procedure.

To support these hypothesis, we performed a face identification experiment that accounts for keypoints clustering in different groups according to the coherence of their descriptors (clustering for $\rho = 2$ is considered). This experiment has been performed on the BU-3DFE database following the same settings discussed in Sect. 2.3. Results for the rank-1 recognition rate (*rank-1* RR) and the related number of keypoint matches (*#matches*) obtained for different groups of clusters and their combinations are reported in Table 3.

It can be observed that clusters in group-1 provide robust recognition. Their combination with the match of isolated keypoints and clusters in group-2 permits to further increase the accuracy of recognition. Instead, considering all the keypoints in face comparison without clustering (row "all keypoints" in the Table) results in a reduction of the accuracy

Table 3 Rank-1 recognition rate (RR) and number of keypoint matches using different groups of keypoints

Cluster group	rank-1 RR (#matches)
group-1	83.1 % (49)
group-1 & isolated keypoints	85.0 % (58)
group-1, 2 & isolated keypoints	88.2 % (79)
All keypoints	85.4 % (133)

with a considerable increase in the number of matches. These results are in agreement with the hypothesis that clustering keypoints based on their spatial proximity and the coherence of the descriptors can enhance the recognition performance both in terms of accuracy and computational cost. Results also support the conclusion that considering only clusters of *group-1* and *group-2* plus isolated keypoints has a positive effect on the recognition. According to these results, in the experiments of Sect. 4, we will consider only clusters of keypoints that fall in *group-1*, *group-2*, and isolated keypoints.

3.3 Feature selection of local descriptors

Local descriptors used to represent the face at 3D keypoints are typically in the form of feature vectors (i.e., histograms) with high dimensionality [21,24,32]. This, combined with the combinatorial number of keypoint correspondences to be computed in the comparison of two faces, can result in a computational cost that does not scale well with large face galleries. In the proposed framework, we considered GHs of 80 bins per ring (see Sect. 2.2), with an overall dimensionality of the mr-GHs that depends on the number of rings used at the local scale (see the analysis of stability in Sect. 3.1).

To reduce the overall number of features used in the comparison of two mr-GH, we propose to use a *feature selection* analysis. This required us to cast the keypoints matching into a classification scenario. The idea we follow here is to select features that are *maximally relevant* and *minimally redundant* in the match of inlier keypoints (i.e., keypoints resulting after RANSAC rejection). To this end, we considered the matches between all the scans in a subset of the BU-3DFE dataset, and recorded the individual distance components in the match of every pairs of keypoints (using the Bhattacharyya distance). In so doing, inlier correspondences between keypoints are marked by a positive label, whereas outlier correspondences are considered as examples with a negative label. These training data are used as input to the *minimal-redundancy maximal-relevance* (mRMR) feature selection model [27]. For a given classification task, the aim of the mRMR algorithm is to select a subset of features by taking into account the ability of features to identify the classification label, as well as the redundancy among the features. These concepts are defined in terms of the mutual information between features. The selected features sorted according to a relevance score are obtained as output of the algorithm.

Results of this analysis are shown in Fig. 11. In (a), the relative relevance of the features is reported as a function of the feature number, where the feature number is obtained by considering the mr-GH descriptor as a one-dimensional array (i.e., the 2D-matrix of each GH is represented as a 1D-array, and these arrays are further concatenated, from the 1st-ring to the last ring). From this plot, it can be observed as several features in the first rings of the mr-GH are less informative than those included in the central and outer ones. Results are shown also in (b), by sorting the features according to their relevance. This provides an indication of the number of features that should be selected if a threshold on their relevance is fixed. For example, 160 features should be selected if a relevance greater than about 35 % is targeted.

The final goal of selecting relevant features is to identify a trade-off between accuracy and effectiveness of the representation. In other terms, the challenge is to identify m out of the n features which yield similar, if not better, accuracies as compared to the case in which all the n features are used. The effect of using only the selected features instead

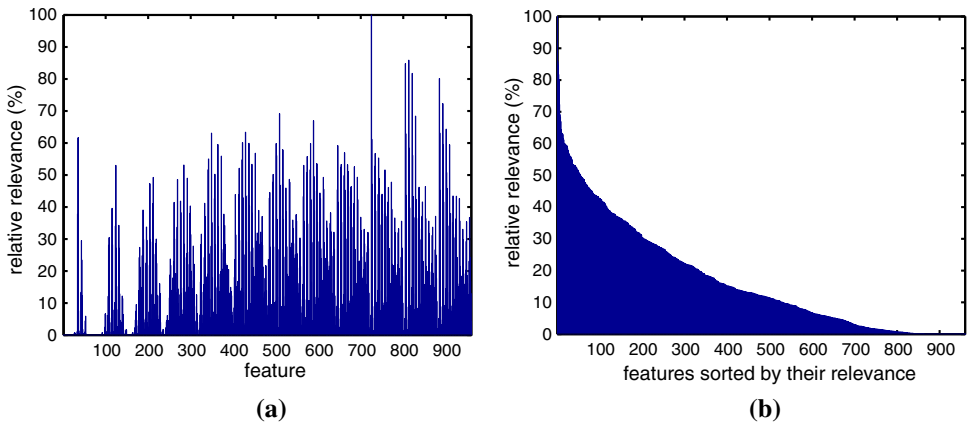


Fig. 11 Relative relevance of the features of the mr-GH descriptor as a function of: **a** The features (identified by their *number*); **b** The features sorted according to their relevance

of the overall descriptor in the comparison of 3D face scans is further investigated in the next section, by considering the impact on the face recognition rate.

4 Experimental results

The proposed approach has been evaluated under different aspects. First, we investigate the number and repeatability of detected keypoints and clusters of keypoints (see Sect. 4.1). Then, we report face identification results on the *Bosphorus* database (see Sect. 4.2), also showing the effect of the criteria proposed in Sect. 3 to improve the stability of keypoints detection and description, and comparing our approach with respect to state of the art solutions. A face authentication experiment following the ROC III protocol on the FRGC v2 dataset and a related comparative analysis is reported in Sect. 4.3. In Sect. 4.4, the effect on the recognition produced by noisy scans acquired with low-resolution cameras is reported. Finally, the possibility to use the detected keypoints as estimators of facial landmarks is investigated in Sect. 4.5.

4.1 Keypoints repeatability

The idea of representing the face by computing local descriptors from a set of detected keypoints relies on the assumption of *intra-subject* keypoints repeatability: keypoints extracted from different facial scans of the same subject are expected to be located approximately in the same positions of the face.

This assumption has been tested on the BU-3DFE database, following the approach proposed in [24]. According to this, we measured the correspondence of the location of keypoints detected in two face scans by performing ICP registration: Two 3D scans of the same subject are automatically registered and the errors between the position of nearest neighbors keypoints (one from each scan) are recorded. Figure 12 shows the results by reporting the cumulative rate of repeatability as a function of increasing distances. The repeatability reaches a value greater than 90 % for frontal faces with neutral and non-neutral expressions at a distance

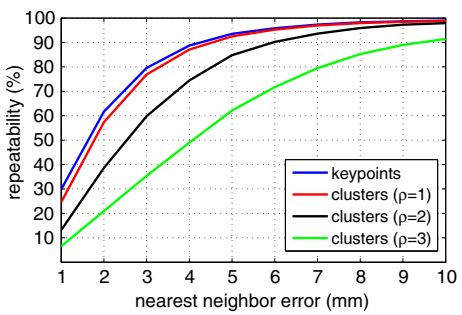


Fig. 12 Repeatability of keypoints

Table 4 Average number of keypoints and clusters of keypoints for different classes of scans in the BU-3DFE database

class (#scans)	#keypoints	#clusters (for $\rho =$)		
		1	2	3
neutral (80)	219	197	112	52
expressive (1,920)	269	222	137	77
total (2,000)	267	221	136	76

error of 5 mm (267 keypoints detected per scan on average). In the same figure, the results obtained by considering the repeatability of the center of the keypoint clusters are also reported, for $\rho = 1, 2, 3$. It can be observed that the repeatability of the cluster centers is lower than that obtained for the keypoints, This can be motivated with the lower number of cluster centers compared to the number of keypoints. However, this reduction is lower than the difference between the number of keypoints and clusters: For example, with $\rho = 2$ at distance 5 mm, the reduction of the repeatability is less than 10 %, whereas the number of clusters is about 50 % than the original number of keypoints (see also the number of keypoints and clusters reported in Table 4). The net result of clustering keypoints and replacing them with cluster centers is the possibility to achieve a robust matching using a lower number of correspondences.

The average number of detected keypoints is also reported in Table 4. It can be observed as non-neutral expressions slightly affect the number of detected keypoints, which remains comparable to that obtained for neutral scans. In general, the number of detected keypoints is quite large. This is in accordance with the results reported in the recent survey on 3D keypoint detectors by Tombari et al. [33], where it is mentioned that meshDOG tends to extract a high number of keypoints, that accumulate around areas characterized by high local curvature. Similar results are also obtained with other 3D keypoint detectors, like that defined by Mian et al. [24], and the meshSIFT [32] (in both the cases, hundreds of keypoints are detected).

In Table 4, the number of clusters of keypoint (*#clusters*) obtained using the solution proposed in Sect. 3.2 for $\rho = 1, 2, 3$ is also reported. A reduction of about 50 % in the number of keypoints is observed using clustering for $\rho = 2$ (compare also with the *keypoints reduction* α in Fig. 8a). This effect, combined with the high repeatability, permitted us to increase the overall face identification accuracy reducing, at the same time, the computational cost in comparing two face scans.

4.2 Face identification on the *Bosphorus* database

Recognition experiments have been performed on the *Bosphorus* database. This dataset has been collected at

Boğaziçi University and released in 2008 [31]. It consists of 3D facial scans and images of 105 subjects acquired under different expressions, various poses and occlusion conditions. Occlusions are given by hair, eyeglasses or predefined hand gestures covering one eye or the mouth. Many of the male subjects have also beard and moustache. The majority of the subjects are Caucasian aged between 25 and 35, with a total of 60 males and 45 females. The database includes a total of 4,666 face scans, with the subjects categorized as follows:

- 34 subjects with up to 31 scans (including 10 expressions, 13 poses, 4 occlusions and 4 neutral);
- 71 subjects with up to 54 different face scans. Each scan is intended to cover one pose and/or one expression type, and most of the subjects have only one neutral face, though some of them have two. Totally, there are 34 expressions, 13 poses, 4 occlusions and one or two neutral faces. In this set, 29 subjects are professional actors/actresses, capable of more realistic and pronounced expressions.

The variability of the scans in terms of subjects' pose, expressions and occlusions motivated us to use this dataset in the experiments. In addition, this dataset has been used by several state of the art solutions for 3D face recognition. In the experiments, we used the same protocol proposed in [21], [32] and [6], thus allowing a direct comparison of the results. For each subject, the first neutral scan was included in the gallery, whereas the probe scans have been organized in different categories as reported in Table 5 (the number of probes per class is also indicated).

The first class groups probes according to their facial expression, distinguishing between neutral probes and expressive probes, plus some not-classified probes. Probes where subjects exhibit face action units (FAU) are accounted in the second class, by considering scans with lower FAU (LFAU), upper FAU (UFAU), and combined action unit (CAU). Finally, the last class reports probes with missing parts due to yaw rotation (YR), pitch rotation (PR) and cross rotation (CR), plus probes with Occlusions (O). For methods in [21,32] and [6], the rank-1 RR is reported as appears in the respective publications. Results of our approach are reported using the optimizations proposed in Sect. 3.

4.2.1 Computational cost

The proposed solution is capable of scoring the same performance of state of the art approaches, reducing at the same time the computational cost. This latter aspect is evidenced in Table 6, where the computational cost in comparing two face scans is approximated by the overall number of distances computed between bins of the local histogram descriptors (last row of the Table). In turn, this number is obtained as product of the average number of matches between keypoints (*matches*)—quadratic in the number of keypoints—and the number of bins of the local descriptors (*bins per descriptor*). For our approach, the number of bins of the local descriptor is considered on average, since it can change from keypoint-to-keypoint depending on the extent of the selected optimal scale, as discussed in Sect. 3.1 (here, the features of the descriptor with relevance greater than 35 % are selected). In our case, we have considered keypoints clustering with $\rho = 2$ and, according to the results of Sect. 3.2, just clusters in group-1, group-2 and isolated keypoints are retained (the sum of these two is reported in the *keypoints* row of the Table). It can be noted as the optimizations proposed in this work combined with the meshDOG/GH are capable of reducing the computational cost to about 1/25 of the values reported in the other cases, without compromising the identification accuracy.

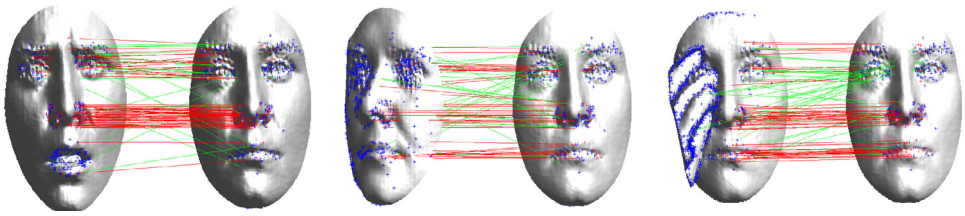
Table 5 *Bosphorus* DB: Rank-1 RR for different probe categories

Probes (#)	Li et al. [21] (%)	Smeets et al. [32] (%)	Berretti et al. [6] (%)	This work (%)
Neutral (194)	100.0	–	97.9	98.5
Anger (71)	88.7	–	85.9	88.7
Disgust (69)	76.8	–	81.2	81.2
Fear (70)	92.9	–	90.0	91.4
Happy (106)	95.3	–	92.5	94.3
Sad (66)	95.5	–	93.9	95.5
Surprise (71)	98.6	–	91.5	94.4
other (18)	–	–	100.0	94.4
LFAU (1,549)	97.2	–	96.5	97.5
UFAU (432)	99.1	–	98.4	99.1
CAU (169)	98.8	–	95.6	96.4
YR (735)	78.0	–	81.6	82.6
PR (419)	98.8	–	98.3	98.8
CR (211)	94.3	–	93.4	95.3
O (381)	99.2	–	93.2	95.8
All (4,561)	94.1	93.7	93.4	94.5

Our approach is compared with the work of Li et al. [21], Smeets et al. [32], and Berretti et al. [6]

Table 6 *Bosphorus* DB: Computational cost in comparing two face scans for our method, and the works in [21], [32] and [6]

Number of	Li et al. [21]	Smeets et al. [32]	Berretti et al. [6]	This work
bins per descriptor	216	288	640	160
keypoints	648	560	377	145
matches ($\cdot 10^3$)	419.9	313.6	142.1	21.0
bin distances ($\cdot 10^6$)	90.7	90.3	90.9	3.4



(a) expression **(b)** missing parts **(c)** occlusion

Fig. 13 Comparison between a gallery scan, and scans of the same subject with: **a** expression, **b** missing parts, **c** occlusion

4.2.2 *Robustness to expressions, missing parts and occlusions*

In general, expressions, missing parts and occlusions of the face are the main factors that impair 3D face recognition methods. These factors act with different modalities on our approach.

Large facial expressions modify locally the 3D shape of the face. This can alter locally the position where keypoints are detected, thus reducing their repeatability. Local descriptors at the keypoints in the deformed region of the face can also change, thus making more difficult to find robust correspondences between keypoints in expressive probe and neutral gallery scans. Robustness of the approach to expressions mainly derives from the capability to match the pairs of keypoints that are least or not at all modified by expressions. In this respect, the clustering analysis reported in Sect. 3.2 is important to reduce the number of instable keypoints. An example of positive match in the case of facial expressions is shown in Fig. 13a. It can be observed as the largest number of matches is located in the nose region, which is less affected by the expression (i.e., fear).

Missing parts of the face affect the approach by reducing the number of detected keypoints (i.e., the surface where they are detected is reduced). In this case, robustness of the match derives from the fact that keypoints detected on the remaining part of probe scans can still correctly match with those in gallery scans. Figure 13b shows the case of a correct match of a probe with part of the right side of the face missing.

In the case large portions of the face are occluded (by hair, glasses, scarf, hand, etc.), keypoints are detected also in the occluded regions. The local descriptors at these keypoints are likely to not match with the keypoints detected in the corresponding non-occluded regions of gallery scans due to the modification that occlusions determine in the 3D surface. Ultimately, this reduces the number of inlier correspondences between probe and gallery scans. However, correspondence between keypoints located in non occluded parts of the face can still be sufficient to grant a correct match. Figure 13c reports the case of a hand covering the right eye and cheek of the face. The good match between keypoints located in the regions of the face that are not occluded can be appreciated.

All in all, expressions and occlusions act mainly on the repeatability of several keypoints of the face and their local descriptors; missing parts, instead, reduce the number of keypoints. The above considerations are supported quantitatively in Table 7, where the number of valid matches in the different cases is reported. In particular, the column *#keypoints* reports the overall number of keypoints (i.e., cluster centers and isolated keypoints) which is used to describe scans of different types, whereas the columns *overall* and *inlier* refer, respectively, to the initial number of keypoint correspondences, and to the number of keypoints matching after outliers rejection using RANSAC. It can be observed that the highest number of keypoint matches is obtained for frontal neutral and expressive scans. In the case of occlusions, the number of matches is reduced, since the keypoints that are detected on the occluded part do not match. Scans with missing parts are the most critical (especially for rotations of more than 45°), since the reduced surface determines a significant decrease in the number of keypoints and their matches.

Table 7 *Bosphorus* DB: Average number of keypoints (#keypoints) and keypoints matching (#keypoints matching) for scans of different categories

Type	#keypoints	#keypoints matching	
		Overall	Inlier
neutral	158	117	93
expressions	165	108	81
occlusions	174	101	60
missing parts	97	68	41

4.3 Face verification on the FRGC dataset

The FRGC v2 dataset includes 3D face scans of 466 subjects partitioned in the *Fall2003* and *Spring2004* sets, respectively, with 1,893 and 2,114 scans (4,007 scans in total). Face scans are given as matrices of points of size 480×640, with a binary mask indicating the valid points of the face. Due to different distances of the subjects from the sensor during acquisition, the actual number of points representing a face can vary. Individuals have been acquired with frontal view from the shoulder level, with very small pose variations. About 60 % of the faces have neutral expression, and the others show expressions of disgust, happiness, sadness,

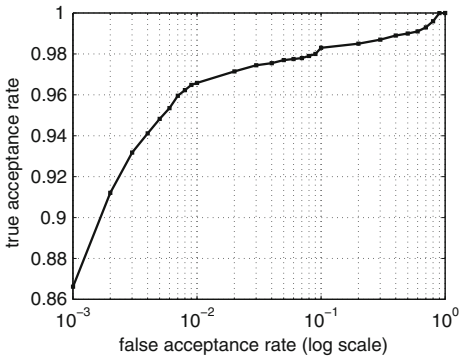


Fig. 14 FRGC v2: ROC III experiment

Table 8 FRGC v2: comparative analysis on the ROC III experiment

Approach	TAR @0.001 FAR (%)
Kakadiaris et al. [18]	97.0
Faltemier et al. [13]	94.8
Al-Olsaimi et al. [1]	94.1
Wang et al. [34]	98.0
Drira et al. [11]	97.1
Lei et al. [19]	96.7
Smeets et al. [32]	77.2
Our approach	86.6

and surprise. Some scans include hair [30]. In the following, we report results obtained by our method on the ROC III face verification experiment of the FRGC v2 protocol. In this experiment, the gallery includes the scans of the *Fall2003* set, whereas the probe scans are from the *Spring2004* set. Due to the time lapse between the acquisition of probe and gallery scans, this experiment is regarded as the most difficult one of the FRGC protocol. Verification results are shown in Fig. 14 using the ROC curve.

Table 8 compares results obtained in the ROC III experiment by our approach and other state of the art solutions (the *True Acceptance Rate* at 0.1 % *False Acceptance Rate*, TAR @0.001 FAR is reported). In this experiment, keypoints-based methods, like ours and that in [32], provide lower accuracy than several other solutions. As also noted in [32], in part this derives from the fact that measuring the similarity based on the number of keypoints matching makes difficult to find a threshold on this number capable of discriminating between matches of same and different subjects. As a consequence, these solutions are more suited to face identification than to face verification. An additional motivation for the performance drop of keypoints-based solutions is that many of the works in the Table have been developed mainly to address expression variations in frontal scans, rather than occlusions and missing parts, and so can better adapt to the FRGC frontal neutral and expressive scans.

4.4 Robustness to noisy data

The proposed approach targets face recognition from high-resolution 3D scans. Since keypoints detection relies on curvature computation, large acquisition noise can negatively affect keypoints detection and description. To investigate this aspect, we performed two experiments as reported in the following.

4.4.1 Synthetic noise on high-resolution scans

This experiment aims to investigate the effect of adding synthetic noise to high-resolution scans. To this end, we consider high-resolution face meshes and create their noisy counterparts by displacing the position of the vertices along the direction of the surface normal (such displacements are the most destructive for the mesh). This operation is repeated 10 times, each time adding noise of increasing maximum magnitude to the original high-resolution scan (i.e., the maximum noise magnitude varies from 1 to 10 mm). On these noisy scans, keypoints detection and description are performed. Figure 15 shows some noisy scans and the corresponding detected keypoints for a given subject.

Then, we evaluated keypoints repeatability between noisy and original scans at varying levels of noise. Results are reported in Fig. 16. It can be observed that keypoints repeatability does not vary significantly for the first level of noise and decreases to about 70 % (at distance 6 mm) at level 3 (i.e., maximum noise magnitude of 3 mm). This level of noise is typically greater than the noise that can be originated by high-resolution acquisition devices of common use. Larger falls of the repeatability are observed for levels of the noise that are destructive for the mesh (i.e., level 5 or greater, as can be also appreciated in Fig. 15d).

4.4.2 Low-resolution noisy scans

In this experiment, we evaluate the combined effect of low-resolution scans and noise by considering real probes acquired by dynamic 3D cameras, like Kinect. The idea here is to model a realistic scenario, where gallery scans are acquired off-line with high-resolution scanners, whereas probes are acquired on-line using low-resolution sensors. To test this application context, we used the *The Florence Superface* (UF-S) dataset [4] that, to the best of our knowledge, is the only dataset which includes both low- and high-resolution scans of the same subjects (there are a few other datasets for face analysis from consumer cameras, like the *EURECOM Kinect Face* dataset [17], or the *The 3D Mask Attack* database [12], but they do not include high- and low-resolution scans of the same subjects). The version 1.0 of the UF-S dataset includes 20 subjects, with the following data captured in the same session: two 3D high-resolution face scan with

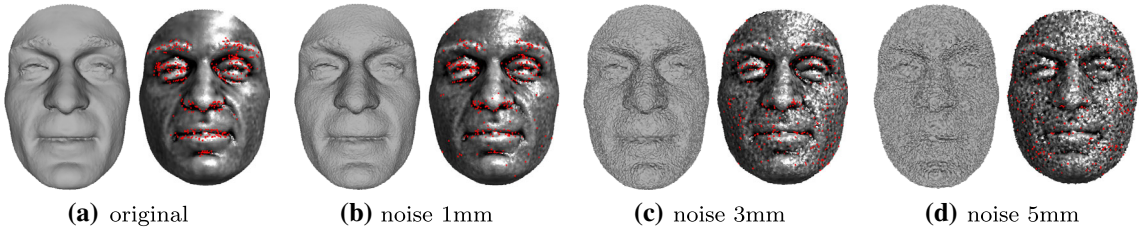


Fig. 15 Effect of adding noise of increasing magnitude in the direction of the normal to the original mesh in (**a**). From **b**–**d** face scans obtained by adding noise of maximum magnitude, respectively, 1, 3 and 5 mm are reported (in each case, the noisy scan and the corresponding detected keypoints are reported)

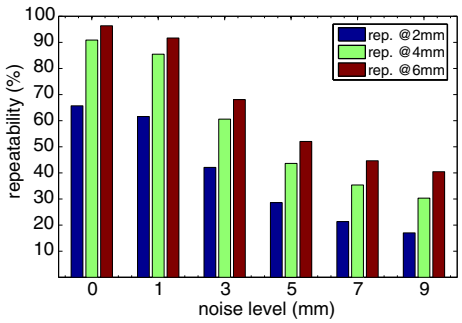


Fig. 16 Keypoints repeatability at different distances as a function of the noise level varying from 0 (no noise) to 9 mm

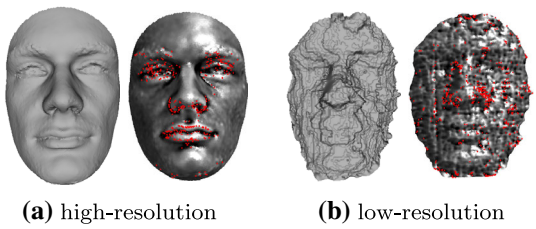


Fig. 17 Florence Superface: **a** High-resolution scan; **b** low resolution acquisition of the same subject in (**a**). In both (**a**) and (**b**), the scans with the detected keypoints highlighted in *red* are also shown

about 40,000 vertices and 80,000 facets, acquired with the *3dMD* scanner; A video sequence acquired with the *Kinect* camera, with the person sitting in front of the sensor at an approximate distance of 80 cm. In this latter acquisition, the subject is also asked to rotate the head around the yaw axis up to an angle of about 50°–60°, so that both the left and right sides of the face are exposed to the camera during acquisition. This results in video sequences lasting approximately 10–15 s, at 30 fps.

Developing a complete method capable of working with Kinect data is not the focus of this work. Rather, using the Kinect frames, we want to show the effect that data with low resolution and large noise can have on 3D keypoints detection and description and on the recognition process. To this end, we defined a pilot recognition experiment, where the 20 high-resolution scans of the UF-S are included in the gallery, and 40 frames of the Kinect sequences in frontal position (2 frames per subject) are considered as low-resolution probes. Examples of high- and low-resolution scans are shown in Fig. 17a and b, respectively. In both the cases, the detected keypoints are also shown. Using this dataset, a rank-1 recognition rate of 57.5 % has been obtained (compared to a 100 % obtained performing the same experiment with high-resolution scans). This evidences the challenge posed by low-resolution noisy scans that can be only partially addressed by our method. An intermediate solution that can alleviate the difficulties arising from consumer cameras can be obtained by deriving super-resolved models from a sequence of low-resolution frames, so as to improve the resolution and reduce, at the same time, the effect of noise [4].

4.5 Keypoints and landmarks of the face

As discussed in Sect. 1, landmarks and keypoints of 3D faces are typically detected following different approaches and with different objectives. However, some keypoints are detected in the close position of landmarks, so that it is possible to investigate the correspondence between keypoints and landmarks of the face. In so doing, we do not aim to provide a complete method for landmarks detection; rather, we want to show further potential applications of keypoints detection on 3D face scans. In this experiment, we used the Bosphorus dataset that provides the 2D and 3D coordinates of up to 24 labelled facial landmarks. These are manually marked on 2D color images, provided that they are visible in the given 3D scan, and the 3D coordinates were then calculated using the 3D-2D correspondences. In our case, we concentrate on nine landmarks (seven of these are also considered in the state of the art work on 3D landmarks detection in [28], and commonly considered as sufficient for many 3D face applications), namely: *Outer/inner, left/right eye corners* (7,8,9,10); *left/right mouth corners* (16,18); *left/right nose peaks* (13,15); *nose tip* (14)—the names and numbers of the landmarks reported in the Bosphorus dataset are used. In our experiment, we considered the location of the landmarks provided in the dataset as ground truth and verified the

Table 9 *Bosphorus* DB: The mean and the standard deviation of the *the absolute distance error* (ADE) and the *detection success rate* (at 10*mm* distance) computed for 9 landmarks of the face

Landmark name (#scans)	ADE (mm)		
	Mean	SD	$\leq$ 10 mm (%)
outer left eye (4,221)	3.04	2.91	97.2
inner left eye (4,146)	3.39	2.92	97.0
outer right eye (4,458)	3.59	3.14	95.2
inner right eye (4,360)	2.68	2.03	98.9
left mouth corner (4,317)	3.89	3.87	93.3
right mouth corner (4,429)	3.42	3.21	95.4
left nose peak (4,113)	5.58	3.89	92.3
right nose peak (4,454)	3.87	3.07	96.3
nose tip (4,662)	8.51	5.37	71.3

The first column reports the landmarks' name and the number of scans for which the landmark is annotated in the ground truth

distance at which a keypoint is detected. Similar to [28], two error measures are used: *Absolute distance error* (ADE), that is the Euclidean distance in millimeters between the position of a keypoint and the manually annotated landmark, which is considered ground truth; *Detection success rate*, which represents the percentage of successful detections of a landmark over the test database considering a distance threshold of 10 mm.

Results are listed in Table 9. It can be observed that, apart for the nose tip, which scores a detection success rate at a distance of 10 mm of about 71 %, for all the other landmarks, this value is greater than 92 % with a maximum of about 99 % for the *inner right eye*. Values of the ADE mean and standard deviation are also small, confirming the possibility to use the proposed 3D keypoints detector as a preliminary step to locate facial landmarks.

5 Conclusions and future work

Face recognition based on the idea of capturing local information around a set of keypoints directly detected in 3D is a promising solution, especially in the case of occlusions or missing parts. However, approaches developed so far that use such framework just proposed basic solutions that do not consider any optimization capable of enabling a better selection of the keypoints and a more effective description of the surface at keypoints' neighborhood. Based on these premises, in this work we have proposed an original analysis that permits an improved stability of the keypoints detection and description. Remarkably, the proposed methods are of general applicability. For a concrete evaluation, they have been applied to a specific approach that includes meshDOG keypoints detector and local GH descriptor. In summary, the approach proposed in this work presents some new solutions in the perspective to make 3D face recognition deployable in real application contexts: The approach is fully-3D, and does not require any costly pose normalization or alignment; the meshDOG keypoints combined with the GHs provide a good solution to achieve robustness to expression changes and to occlusions/missing parts of the face; the proposed methods for selecting stable keypoints and for their clustering, combined with the selection of optimal features of the local descriptors, have permitted an increased accuracy as well as a lower computational cost of the recognition.

In perspective, the proposed framework could be easily adapted to include appearance of the face, so as to define a multi-modal solution that combines together, in the function used for meshDOG detection, 2D and 3D data. Promising appears also the use of 3D keypoints to locate landmarks of the face. Additional investigations are instead required to make this approach effective in the case of data acquired by low-resolution depth cameras.

References

1. Al-Osaimi, F.R., Bennamoun, M., Mian, A.: An expression deformation approach to non-rigid 3D face recognition. Int. J. Comput. Vis. **81**(3), 302–316 (2009) NULL
2. Ashbrook, A., Fisher, R., Robertson, C., Werghi, N.: Finding surface correspondance for object recognition and registration using pairwise geometric histograms. In: Proceedings of European Conference on Computer Vision, pp. 674–686. Friburg (1998)
3. Berretti, S., Del Bimbo, A., Pala, P.: 3D face recognition using isogeodesic stripes. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 2162–2177 (2010)
4. Berretti, S., Del Bimbo, A., Pala, P.: Superfaces: A super-resolution model for 3D faces. In: Proceedings Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment. Firenze (2012)
5. Berretti, S., Del Bimbo, A., Pala, P.: Sparse matching of salient facial curves for recognition of 3D faces with missing parts. IEEE Trans. Inf. Forensics Secur. **8**(2), 374–389 (2013)
6. Berretti, S., Werghi, N., Del Bimbo, A., Pala, P.: Matching 3D face scans using interest points and local histogram descriptors. Comput. Graph. **37**(6), 509–525 (2013)
7. Bowyer, K.W., Chang, K.I., Flynn, P.J.: A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. Comput. Vis. Image Underst. **101**(1), 1–15 (2006)
8. Boyer, E., Bronstein, A.M., Bronstein, M.M., Bustos, B., Darom, T., Horaud, R., Hotz, I., Keller, Y., Keustermans, J., Kovnatsky, A., Litman, R., Reininghaus, J., Sipiran, I., Smeets, D., Suetens, P., Vandermeulen, D., Zaharescu, A., Zobel, V.: SHREC 2011: Robust feature detection and description benchmark. In: Proceedings of Eurographics Workshop on 3D Object Retrieval (3DOR 2011). Llandudno (2011)
9. De Carlo, D., Metaxas, D., Stone, M.: An anthropometric face model using variational techniques. In: Proceedings of ACM SIGGRAPH, pp. 67–74. Orlando (1998)
10. Dorkó, G., Schmid, C.: Maximally stable local description for scale selection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) Computer Vision ECCV 2006. Lecture Notes in Computer Science, vol. 3954, pp. 504–516. Springer, Berlin (2006)

11. Drira, H., Amor, B.B., Srivastava, A., Daoudi, M., Slama, R.: 3D face recognition under expressions, occlusions and pose variations. IEEE Trans. Pattern Anal. Mach. Intell. **35**(9), 2270–2283 (2013)
12. Erdogmus, N., Marcel, S.: Spoofing in 2D face recognition with 3D masks and anti-spoofing with kinect. In: Proceedings of IEEE International Conference on Biometrics: Theory, Applications and Systems. Washington DC (2013)
13. Faltemier, T.C., Bowyer, K.W., Flynn, P.J.: A region ensemble for 3D face recognition. IEEE Trans. Inf. Forensics Secur. **3**(1), 62–73 (2008)
14. Farkas, L.G.: Anthropometry of the Head and Face. Raven Press, New York (1994)
15. Goswami, G., Bharadwaj, S., Vatsa, M., Singh, R.: On RGB-D face recognition using Kinect. In: Proceedings of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). Washington DC (2013)
16. Huang, D., Ardabilian, M., Wang, Y., Chen, L.: 3D face recognition using eLBP-based facial representation and local feature hybrid matching. IEEE Trans. Inf. Forensics Secur. **7**(5), 1551–1565 (2012)
17. 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: Proceedings of ACCV Workshop on Computer Vision with Local Binary Pattern Variants. Daejeon (2012)
18. Kakadiaris, I.A., Passalis, G., Toderici, G., Murtuza, N., Lu, Y., Karampatziakis, N., Theoharis, T.: Three-dimensional face recognition in the presence of facial expressions: an annotated deformable approach. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 640–649 (2007)
19. Lei, Y., Bennamoun, M., Guo, M.H.Y.: An efficient 3D face recognition approach using local geometrical signatures. Pattern Recognit. **47**(2), 509–524 (2014)
20. Li, B.Y.L., Mian, A.S., Liu, W., Krishna, A.: Using kinect for face recognition under varying poses, expressions, illumination and disguise. In: Proceedings of IEEE Workshop on Applications of Computer Vision, pp. 186–192. Clearwater (2013)
21. Li, H., Huang, D., Lemaire, P., Morvan, J.M., Chen, L.: Expression robust 3D face recognition via mesh-based histograms of multiple order surface differential quantities. In: Proceedings of IEEE International Conference on Image Processing, pp. 3053–3056 (2011)
22. Lowe, D.: Distinctive image features from scale-invariant key points. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
23. Maes, C., Fabry, T., Keustermans, J., Smeets, D., Suetens, P., Vandermeulen, D.: Feature detection on 3D face surfaces for pose normalisation and recognition. In: IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6. Washington DC (2010)
24. Mian, A.S., Bennamoun, M., Owens, R.: Keypoint detection and local feature matching for textured 3D face recognition. Int. J. Comput. Vis. **79**(1), 1–12 (2008)
25. Min, R., Choi, J., Medioni, G., Dugelay, J.L.: Real-time 3D face identification from a depth camera. In: Proceedings of International Conference on Pattern Recognition, pp. 1739–1742. Tsukuba (2012)
26. Passalis, G., Perakis, P., Theoharis, T., Kakadiaris, I.A.: Using facial symmetry to handle pose variations in real-world 3D face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 1938–1951 (2011)
27. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency max-relevance and min-redundancy. EEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
28. Perakis, P., Passalis, G., Theoharis, T., Kakadiaris, I.A.: 3D facial landmark detection under large yaw and expression variations. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1552–1564 (2013)
29. Peyre, G.: Toolbox graph. In: MATLAB Central File Exchange Select (2009)
30. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Proceedings of IEEE Workshop on Face Recognition Grand Challenge Experiments, pp. 947–954. San Diego (2005)
31. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3D face analysis. In: Proceedings of COST 2101 Workshop on Biometrics and Identity Management (2008)
32. Smeets, D., Keustermans, J., Vandermeulen, D., Suetens, P.: meshSIFT: Local surface features for 3D face recognition under expression variations and partial data. Comput. Vis. Image Underst. **117**(2), 158–169 (2013)
33. Tombari, F., Salti, S., Di Stefano, L.: Performance evaluation of 3D keypoint detectors. Int. J. Comput. Vis. **102**(2–3), 198–220 (2013)
34. Wang, Y., Liu, J., Tang, X.: Robust 3D face recognition by local shape difference boosting. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 1858–1870 (2010)
35. Werghi, N., Rahayem, M., Kjellander, J.: An ordered topological representation of 3D triangular mesh facial surface: concept and applications. EURASIP J. Adv. Signal Process. **2012**(144), 1–20 (2012)
36. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3D facial expression database for facial behavior research. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp. 211–216. Southampton (2006)
37. Zaharescu, A., Boyer, E., Varanasi, K., Horaud, R.: Surface feature detection and description with applications to mesh matching. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 373–380. Miami Beach (2009)
38. Zuliani, M., Kenney, C.S., Manjunath, B.S.: The multiransac algorithm and its application to detect planar homographies. In: Proceedings of IEEE International Conference on Image Processing, pp. 153–156. Genoa (2005)

Stefano Berretti received the PhD in Information and Telecommunications Engineering in 2001 from the University of Firenze, Italy. Currently, he is an associate professor at the Department of Information Engineering of the University of Firenze, Italy, and at the Media Integration and Communication Center of the same University. His current research interests are mainly focused on content modeling, retrieval, and indexing of image and 3D object databases. Recent researches have addressed 3D object retrieval and partitioning, 3D face recognition, 3D facial expression recognition. He has been visiting researcher at the Indian Institute of Technology (IIT), in Mumbai, India (2000), and visiting professor at the Institute TELECOM, TELECOM Lille 1, in Lille, France (2009), and at the Khalifa University of Science, Technology and Research, Sharjah, UAE (2013). Stefano Berretti is author of more than 100 publications appeared in conference proceedings and international journals in the area of pattern recognition, computer vision and multimedia. He is in the program committee of several international conferences and serves as a frequent reviewer of many international journals. He has been co-chair of the Fifth Workshop on

Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA'12), held on October 7, 2012, in conjunction with ECCV 2012, Firenze, Italy.

Naoufel Werghi received PhD in Computer Vision from the University of Strasbourg. He has been a research fellow at the Division of Informatics in the University of Edinburgh, Lecturer at Department of Computer Sciences in the University of Glasgow. Currently, he is an associate professor at the Electrical and Computer Engineering Department in Khalifa University, UAE. His main research area is image analysis and interpretation where he has been leading several funded projects in the areas of biometrics, medical imaging, geometrical reverse engineering, and intelligent systems. He published more than 70 journal and conference papers.

Alberto del Bimbo is full professor of Computer Engineering, Director of the Master in Multimedia, and Director of the Media Integration and Communication Center at the University of Florence. He was the Deputy Rector for Research and Innovation Transfer of 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 is the author of the monography "Visual Information Retrieval." From 1996 to 2000, he was the President of the IAPR Italian Chapter and from 1998 to 2000, 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, IEEE ICMCS'99, the International Conference on Multimedia Computing and Systems and Program Co-Chair of ACM Multimedia 2008. He is the General Co-Chair of ACM Multimedia 2010 and of ECCV 2012, the European Conference on Computer Vision. He is IAPR Fellow and 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 Associate Editor of Pattern Recognition, IEEE Transactions on Multimedia and IEEE Transactions on Pattern Analysis and Machine Intelligence.

Pietro Pala received the PhD in Information and Telecommunications Engineering in 1997 at the University of Firenze, Italy. Pietro Pala is currently an associate professor and the President of the committee deputed to the evaluation of quality, according to the CRUI 2011 model, for the course of Informatics Engineering. 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, segmentation of 3D models. Recently, the research activity focused on the study of biometric models for person recognition based on 3D facial scans. Pietro Pala serves as editor for Multimedia Systems and as reviewer for many leading scientific journals including IEEE Trans. on Pattern Analysis and Machine Intelligence, IEEE Trans. on Multimedia, ACM Trans. on Multimedia Computing Communications and Applications and Pattern Recognition.