Computers & Graphics 37 (2013) 509–525

Contents lists available at SciVerse ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on 3D Object Retrieval

Matching 3D face scans using interest points and local histogram descriptors ☆

Stefano Berretti [a,*], Naoufel Werghi [b], Alberto del Bimbo [a], Pietro Pala [a]

[a] Department of Information Engineering, University of Firenze, Italy
[b] Khalifa University of Science Technology & Research, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form
29 March 2013
Accepted 1 April 2013
Available online 1 May 2013

Keywords:
3D face recognition
3D keypoints
3D local descriptors

ABSTRACT

In this work, we propose and experiment an original solution to 3D face recognition that supports face matching also in the case of probe scans with missing parts. In the proposed approach, distinguishing traits of the face are captured by first extracting 3D keypoints of the scan and then measuring how the face surface changes in the keypoints neighborhood using local shape descriptors. In particular: 3D keypoints detection relies on the adaptation to the case of 3D faces of the meshDOG algorithm that has been demonstrated to be effective for 3D keypoints extraction from generic objects; as 3D local descriptors we used the HOG descriptor and also proposed two alternative solutions that develop, respectively, on the *histogram of orientations* and the *geometric histogram* descriptors. Face similarity is evaluated by comparing local shape descriptors across inlier pairs of matching keypoints between probe and gallery scans. The face recognition accuracy of the approach has been first experimented on the difficult probes included in the new 2D/3D Florence face dataset that has been recently collected and released at the University of Firenze, and on the Binghamton University 3D facial expression dataset. Then, a comprehensive comparative evaluation has been performed on the Bosphorus, Gavab and UND/FRGC v2.0 databases, where competitive results with respect to existing solutions for 3D face biometrics have been obtained.

1. Introduction

The humans' cognitive system has a peculiar attitude in recognizing faces with high accuracy, at least for familiar people in favorable viewing conditions (i.e., good illumination, small occlusions, etc.). Automatic identity recognition performed by machines has entered the scene some decades ago with the aim to extend the human capabilities by covering different and more general contexts. In particular, face has affirmed itself as one of the most important biometric trait due to the fact that images or videos of the face are collectable in an easy and non-intrusive way, whereas other biometrics, such as fingerprints or iris scans are impractical to implement in many scenarios (e.g., in a surveillance setting). Impressively, recent studies report that automatic face recognition can even outperform the human performance in some particular conditions [1]. However, the accuracy of automatic identity recognition based on faces still suffers from many factors, such as pose changes, illumination variations, facial expressions and occlusions.

To solve these problems, face recognition using 3D scans of the face has been recently proposed as an alternative or complementary solution to conventional 2D face recognition approaches using still images or videos, so as to allow accurate face recognition also in real-world applications with unconstrained acquisition. Confirming this recent research trend, several 3D face recognition approaches have been proposed and experimented in the last few years (see the survey in [2], and the literature review in [3–5] for a thorough discussion). However, many of the works appeared in this field, proposed conventional face recognition experiments, where both the probe and gallery scans are assumed to be acquired cooperatively in a controlled environment in which the whole face is precisely captured and represented. These methods mainly focussed on face recognition in the presence of expression variations, reporting very high accuracy on benchmark databases like the *Face Recognition Grand Challenge* (FRGC version 2.0) [6]. Recent studies also exploit ethnicity, gender and age to improve the accuracy of 3D face recognition [7,8]. Solutions enabling face recognition in uncooperative scenarios are now attracting an increasing interest. In such a case, probe scans are acquired in unconstrained conditions that may lead to *missing parts* (non-frontal pose of the face) or to *occlusions* due to hair, glasses, scarves, hand gestures, etc. These difficulties are further sharpened by the recent advent of 4D scanners (3D plus time) [9–11], capable of acquiring temporal sequences of 3D scans. In fact, the dynamics of facial movements captured by these devices can be useful for many applications [12,13], but also increases the acquisition noise and the variability in subjects' pose. In summary, techniques supporting *3D partial face matching* are gaining

☆To comment on this article, please join the discussion on the Collage Authoring Environment Google Group https://groups.google.com/group/collage-authoring-environment.

* Corresponding author. Tel.: +39 55 4796415; fax: +39 55 4796493.
E-mail addresses: berretti@dsi.unifi.it, stefano.berretti@unifi.it (S. Berretti).

0097-8493
http://dx.doi.org/10.1016/j.cag.2013.04.001

importance in making 3D face recognition techniques deployable in more general contexts and, in perspective, in scenarios where 3D dynamic acquisition is performed. However, the research in this context is still preliminary also due to the limited number of face databases that also comprise partial acquisitions of 3D faces [14–16].

1.1. Related work

Below, we review the most recent methods for 3D face recognition, by limiting our analysis to the works that also propose and evaluate solutions supporting partial match of 3D facial scans. In particular, we focus on methods that were also evaluated on scans acquired from non-frontal views of the face for which the recognition problem is further complicated by artifacts that alter the geometry of the acquired 3D surface in correspondence to the borders of the missing regions, rather than to solutions that just cropped 3D full face scans to simulate missing parts. In general, existing solutions can be grouped as *global* and *local*; *Multimodal* approaches that combine together 2D and 3D methods are also possible.

Global 3D face representations for partial face matching have been proposed in a limited number of works. The first solutions appeared in this category used the Iterative Closest Point (ICP) algorithm [17]. The method proposed in [18], was global and multimodal trying to combine 3D shape and 2D texture to perform surface and appearance-based matching. The surface matching component was based on a coarse to fine alignment between a 2.5D probe and a fully 3D face model (obtained by the fusion of five 2.5D scans). In the coarse step, first three manually labeled generic points were used to calculate the rigid transformation that aligns the 2.5D scan with the 3D model, then specific feature points are identified by finding correspondence between shape index values of two scans. These feature points are then used to define a grid of control points around them. In the fine alignment step, a modified ICP algorithm is applied on the grid of control points to refine the alignment between 2.5D probes and 3D models. Good results were reported for neutral, expressive and partial scans of a proprietary database of 200 individuals, though the computational cost does not scale to large datasets. Following a similar idea, 3D face matching between 2.5D probe scans and fully 3D models is proposed in [19]. Also in this case, a coarse alignment is first performed based on the manual labeling of three generic points in the two matching scans, then ICP fine alignment is performed and the registration error is used to evaluate the similarity between the two matching scans. Separate results for scans acquired with moderate expressions, illumination changes and left/right pose variations were reported on a database of 50 subjects. The main limitations of the approach are in the scalability of ICP, and the manual labeling required by the initial coarse alignment. A canonical representation of the face is proposed in [20], where the isometry invariance of the face surface is exploited to manage missing data obtained by randomly removing areas from frontal face scans. However, no side scans were used for recognition. In [21], results on partial face matching removing quadrants of the FRGC v2.0 probes and using face crops around the nose tip are reported. This approach relies on the symmetry of the 3D face scans in order to identify the nose tip and register depth maps so as to derive a Pure Shape Difference Map (PSDM) between pairs of matching scans. Unfortunately, the symmetry hypothesis used for the registration and fiducial points detection is often violated when side views of the face are acquired in uncooperative scenarios. Instead, the experiments are conducted by just removing parts of the face after the preprocessing has been performed on the entire scans. The fact that the same part of the face is removed from both probe and gallery scans in order to generate the PSDM also reduces the concrete applicability of the approach.

The approaches above provide a *global* modeling of both gallery and probe scans, but more successful and scalable solutions use *local* representations of the face. A possible way to solve locally the problem of missing data in 3D face acquisition is to detect the absence of regions of the face and use the existing data to reconstruct the missing parts. The reconstructed scan can then be used as an input to conventional 3D face recognition methods that assume that the entire scan is available. This approach is followed in [22], focusing on face occlusions induced by glasses, scarves, caps, or by the subject's hand. A generic facial model and thresholding on facial surface distances are used to detect occlusions. In this way, the occluded areas are detected and the missing regions are restored using information from the non-occluded parts. However, face recognition accuracy was not evaluated. In [23,24], an inter-pose face recognition solution is proposed which exploits the hypothesis of facial symmetry to recover missing data in facial scans with large pose variations. First, an automatic face landmarks detector is used to identify the pose of the facial scan by marking regions of missing data and roughly registering the facial scan with an Annotated Face Model (AFM) [25]. Then, the AFM is fitted using a deformable model framework that exploits facial symmetry where data are missing. Wavelet coefficients extracted from a geometry image derived from the fitted AFM are used for the match. Experiments have been performed using the *University of Notre Dame* (UND) database [15], with the FRGC v2.0 gallery scans and side scans with 45° and 60° rotation angles respectively as probes. Since it is based on the left/right facial symmetry, this solution can work as long as half of the face with respect to the yaw axis is visible in the scan.

Tackling the problem from an opposite perspective, some methods divide the face into regions and try to restrict the match to uncorrupted parts of the face. Following this idea, the approach in [26] accurately identifies the nose tip in order to extract multiple overlapping regions around it. These regions are matched using the ICP algorithm and the respective scores are combined together in order to evaluate face similarity. This method is extended in [27] by using a set of 38 regions that densely cover the face, and selecting the best-performing subset of 28 regions to perform matching using the ICP algorithm. A recognition experiment accounting for partial match is reported that uses the left and right parts of the FRGC v2.0 probes. However, the experiments only account for the case in which some of the extracted regions are missing, rather than considering the more general case where also parts of the regions can miss. A part-based 3D face recognition method is proposed in [28], which operates in the presence of both expression variations and occlusions. The approach is based on the use of Average Region Models (ARMs) for registration: The facial area is manually divided into several meaningful components, such as eye, mouth, cheek and chin regions, and registration of faces is carried out by separate dense alignment of the regions with respect to the corresponding ARMs. The dissimilarities between gallery and probe scans obtained for individual regions are then combined to determine the final dissimilarity score. Under variations, like those caused by occlusions, the method can determine noisy regions and discard them. The performance of this approach is tested on the *Bosphorus*3D face database [16] that includes facial expressions, pose differences and occlusions. However, a strong limitation of this solution is the use of manually annotated landmarks that are required for both face alignment and regions segmentation. Instead of using extended regions, in [29] the face is represented by a collection of radial curves originating from the nose tip. Face comparison is obtained by elastic matching of the curves. A quality control allows the exclusion of corrupted radial curves from the match, thus enabling the recognition also in the case of missing data. Results of partial matching are given for the side scans of the *Gavab* database [14].

Methods that perform face recognition based on regions, use some landmarks of the face to identify the regions of interest for

the match. However, facial landmarks are difficult to detect when the pose significantly deviates from the frontal one. In addition, since parts of the regions can be missing or occluded, the extraction of effective region descriptors is hindered, so that regions comparison is mostly performed using rigid (ICP) or elastic registration (*deformable models*). Approaches that use keypoints of the face promise to solve some of these limitations. Rather than relying on the detection of specific regions of the face that can fail in the presence of occlusions and missing parts, they assume that detection of keypoints on the face surface and description of these keypoints yield robust yet accurate representation of facial traits, also in the presence of occlusions and missing parts. In doing so, the number of keypoints is supposed to be sufficiently high. In this perspective, the use of keypoints instead of facial landmarks is advantageous. In fact, just few facial landmarks can be accurately detected in an automatic way – from three to ten are at most reported [30] – and detection of a larger number of landmarks is difficult even through partial manual assistance. In the case of partial face scans, up to half of these points are typically not detectable, so that description of such points and of their relationships is of limited effectiveness for face recognition. Differently, a much larger number of keypoints are typically detected – from tens to hundreds of keypoints can be easily derived – and their distribution is rather sparse, not being constrained to specific locations of the face. This makes keypoints more robust than landmarks to missing parts and also allows the extraction of a large number of local descriptors of the face. A first approach that exploits keypoints of the face has been reported in [31], where a 3D keypoints detector and descriptor inspired by the Scale Invariant Feature Transform (SIFT) [32] have been designed. This detector/descriptor has been used to perform 3D face recognition through a multi-modal 2D+3D approach that also uses the SIFT detector/descriptor to index 2D texture face images. However, results do not account for scans with pose variations and missing parts. The 3D keypoints detector defined in [31] was further generalized to the match of generic objects in [33]. Use of keypoints for partial face matching has been recently reported in [34,35]. In this approach, Multi-Scale Local Binary Patterns (MS-LBP) and Shape Index (SI) are applied to face depth images, and the scalar values obtained at each pixel are used to create an MS-LBP map and an SI map. On both these maps, the SIFT detector and descriptor are used to represent local variations of the features extracted from the face. Finally, the matching scheme accounts for local and global face features by combining local matches between SIFT features, with global constraints originated by facial components. Partial face matching results are presented for the FRGC v2.0 scans where parts of the face are masked to simulate missing parts. However, as pointed out by the authors, the approach can deal automatically just with nearly frontal face data as those included in the FRGC v2.0 dataset. In the case of missing parts of the face due to large pose variations the approach is likely to fail. Methods in [36,37] use keypoints detection for the purpose of partial face matching, resulting the best performing approaches in the track on *3D Face Models Retrieval* of the SHREC'11 competition [38]. In particular, in [36] an extension of SIFT and index map based SIFT matching [34] is proposed. First, feature points are detected on each 3D face scan using *mesh* SIFT [39]; then, the quasi-daisy local shape descriptor [40] of each feature point is obtained using multiple order histograms of differential quantities extracted from the surface; Finally, these local descriptors are matched by computing their orientation angles. The number of matched points is used as similarity between two face scans. In [37], first a PCA based shape model is learned by registering a set of training scans to a reference template model (using 12 manually annotated landmarks) and subsequently warping the template on the training scans using a non-rigid registration based on variational implicit functions. The learned model is then fitted to probe and gallery scans to generate model-based descriptions used to evaluate scans similarity. In this approach, *mesh* SIFT is used to detect keypoints whose correspondences in different scans permit to initialize the pose of probe and gallery scans with respect to the model (anyway, a manual initialization is required for about 2.5% of the scans). After pose initialization, the model is fitted following a Bayesian strategy with outliers detection and estimation. The result is an EM alike optimization, where the model updates are alternated with outlier updates, iteratively.

1.2. Contribution and organization

In this work, we propose an original 3D face recognition approach which is also capable to perform recognition in the case parts of the face scans are missing. We rely on the observation that describing the face with local geometric information extracted at the neighbors of keypoints allows partial face matching in which no particular assumption about the number or locations of the keypoints is necessary to perform sparse keypoints matching. In so doing, the size of the support used to compute the local descriptor at keypoint locations becomes crucial: small supports reduce the effectiveness of the descriptor and large supports are more sensible to missing parts that can alter the support itself. In addition, discriminant facial features are not only related to local characteristics of the face surface in the proximity of a set of keypoints, but also to mutual relationships among the position of the keypoints on the face.

Based on these premises, we propose a 3D face description approach that relies on the detection of 3D keypoints on the face surface and the description of the surface in correspondence to these keypoints. In contrast to solutions where keypoints correspond to meaningful face landmarks, such as the *eyebrows*, *eyes*, *nose*, *cheek* and *mouth* [30], we do not exploit any particular assumption about the position of the keypoints on the face surface. Rather, we expect the position of keypoints to be influenced by the specific morphological traits of the face of each subject. In particular, we exploit the assumption of *within subject keypoints repeatability*: the position of the most stable keypoints – detected at the coarsest scales – do not change substantially across facial scans of the same subject. According to this, we propose an adaptation of the meshDOG [41,42] algorithm to the specific case of 3D faces as 3D keypoints detector. In fact, meshDOG has been introduced as 3D extrema detector for the case of generic 3D objects, proving its effectiveness. However, to the best of our knowledge, it has never been applied before to the case of 3D face matching. Then distinguishing traits of a face scan are captured by local descriptors at the detected keypoints. In particular, we experiment the meshHOG descriptor [41], and also propose and experiment two different local descriptors, namely the *histogram of orientations* (SHOT) and the *geometric histogram* (GH), which exploit local properties of the mesh in different ways. We point out that all the processing required to detect keypoints and extract their local descriptors is performed on 3D meshes without requiring any pose normalization or landmark detection. In the comparison of two faces, local descriptors at the 3D keypoints are matched in order to determine the keypoints correspondences. Spatial constraints using RANSAC [43] are also imposed to avoid outlier matches.

Our approach has been experimentally evaluated with a two-fold objective. On the one hand, we verified the accuracy of recognition on two datasets that include probes with extreme variations in terms of facial expressions (*The Binghamton University 3D facial expression dataset* (BU-3DFE) [44]), and probes with up to half of the face missing due to acquisitions with large pose

variations (the *2D/3D Florence Face database* (UF-3D) [45]). On the other, we experimented our solution on three largely used benchmark datasets (namely *Bosphorus*, *Gavab* and *UND/FRGC v2.0*) which allow the comparison of our solution with respect to state of the art approaches.

The contribution of our approach and its novelty over existing solutions using a similar framework, including keypoints extraction, local description and keypoints matching [31,36,39,46], can be summarized as follows:

- *Method*—An original adaptation of the meshDOG detector to the case of face meshes; The adaptation and comparison of three mesh descriptors to the case of 3D faces and their use as local representation at the keypoints; Proposal of the *multi-ring GH* as the local descriptor at the keypoints, and its identification as the most suitable descriptor to be combined with meshDOG keypoints, providing accurate recognition both in the presence of expression variations and large missing parts of the face; A 3D keypoints matching that also encompasses outliers removal using RANSAC.
- *Experiments*—This work contributes an original experimental validation on the new UF-3D face dataset that has never been used before for the purpose of 3D face recognition. Results reported by our work on this dataset can be regarded as a reference evaluation for future works aiming to test 3D face recognition approaches on challenging scans with missing parts; A thorough experimentation on the large and extreme facial expressions included in the BU-3DFE; A comprehensive comparative evaluation that includes the *Bosphorus*, *Gavab* and *UND/FRGC v2.0* datasets.

The remaining content of the paper is organized as follows: In Section 2, we present the adaptation of the meshDOG detector to the case of 3D faces, and we motivate and discuss the relevance of detected keypoints; Local descriptors computed at the keypoints are reported in Section 3; The way local keypoint descriptors are matched in two scans under comparison, so as to permit identity recognition is detailed in Section 4; A thorough experimental validation and comparison are reported in Section 5; Finally, results are discussed and future research directions are outlined in Section 6.

2. 3D keypoints

Several keypoint detectors capable to identify salient points on 3D meshes have been recently proposed. For a thorough comparative evaluation the reader can refer to the recent report at the SHREC'11 contest [47] (track on "robust feature detection and description benchmark") and to the performance evaluation reported in [48]. Among these methods, the meshDOG detector [41,42] has been proved to be superior, in terms of both repeatability of the detection and accuracy of the matching, to other 3D keypoint detectors/descriptors, like the Harris 3D [49], meshSIFT [39,46] and Shape MSER [50] (see the results in [47] for a comparative analysis, and also the comparison provided in [48]). In particular, the meshDOG detector is proposed to perform feature detection, while the meshHOG descriptor is used for the purpose of mesh matching between generic 3D meshes. However, in the work of Zaharescu et al. [41], the 3D keypoints (extrema) were used for matching generic objects, like 3D reconstruction of the human body, reconstructed and synthetic 3D objects, using photometric surface information to extract the object descriptors using meshHOG. To the best of our knowledge, the meshDOG detector has never been used before for the purpose of 3D face analysis. In the following, we present the adaptation of the method so as to make it appropriate for extracting keypoints of 3D face meshes.

2.1. meshDOG of face meshes

The keypoints detection starts by defining and computing a scalar function f on a 3D mesh S. In principle, the function f can be any scalar function $f(v): S \rightarrow R$ that for any vertex $v \in S$ returns a scalar value. This can comprise functions computed according to the chromatic appearance of the mesh surface as well as functions that consider properties of the surface like the mean or Gaussian curvatures. In our case, we used the *mean* curvature at vertex v as value of the function $f(v)$. Though such function is not completely intrinsic, and therefore not completely invariant to local isometric deformations, in practice the keypoints detected using mean curvature turned out to be more stable on 3D face data than keypoints obtained using Gaussian curvature. One motivation for this can be the average operation, which has the advantage to smooth the noise effect that can be present in the computation of principal curvatures. The choice of the mean curvature is also supported in the recent survey on the evaluation of 3D keypoint detectors by Salti et al. [48], where the mean curvature is reported to provide better results than Gaussian curvature when combined with the meshDOG detector. The same conclusion was also reported by the authors of the meshSIFT approach [39,46], where the mean curvature was used in the construction of their scale-space extrema. According to [51], the mean curvature is computed by first rotating the local neighborhood of a vertex so that the normal of the current vertex is aligned with the Z-axis, and the neighborhood can be described by XY only, instead of XYZ. Then, a least-squares quadratic patch is fitted to the local neighborhood of a vertex $h(x,y) = ax^2 + by^2 + cxy + dx + ey + g$, and the eigenvectors and eigenvalues of the Hessian matrix are used to calculate the principal and mean curvature of the vertex.

Once the function f (mean curvature) is computed for every vertex of the mesh, the keypoints selection proceeds by processing the values of the function f through three subsequent steps. In the first step, the extrema of the Laplacian's function (DOG) across scales are found using a one-ring neighborhood of each vertex. Then, the extrema are sorted and thresholded based on a percentage value of the overall number of extrema. Finally, in the third step, only the extrema with some degree of cornerness are retained, thus removing unstable extrema. Details of these steps are given in the following.

Extrema of the scale-space. As first step, a scale-space representation of the scalar function f defined on the mesh is constructed. At every scale, the function f is convolved with a Gaussian kernel (see Eq. (A.3) for the definition of the convolution on the mesh)

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

where σ is the standard deviation of the Gaussian (set equal to $\sigma = 2^{1/3} e_{avg}$ in our experiments, being e_{avg} the average edge length); and, at a vertex v_i, x is the distance between neighboring vertices to the vertex v_i, that is $\|\mathbf{v}_j - \mathbf{v}_i\|$.

The scale-space of f is built incrementally on $N+1$ levels, so that $f_0 = f$, $f_1 = f_0 * g_\sigma$, $f_2 = f_1 * g_\sigma$, …, $f_N = f_{N-1} * g_\sigma$. The *N Difference of Gaussian* (DOG) are then obtained by subtracting adjacent scales, e.g., $DOG_1 = f_1 - f_0$, $DOG_2 = f_2 - f_1$, …, $DOG_N = f_N - f_{N-1}$. In so doing, it is relevant to note that in building the scale space, the geometry of the face does not change, but the different scalar functions f_k and DOG_k defined on the mesh. A total of 96 convolutions (i.e., scales) have been used in our work. Once the scale-space is computed, the feature points are selected as the maxima of the *DOG* across scales. In particular, a vertex is an extremum at a given scale k if its DOG_k

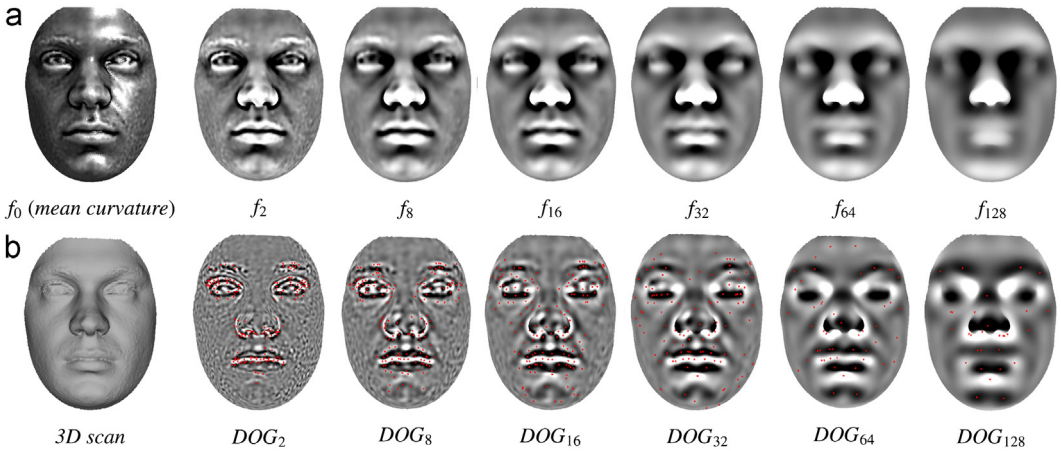


Fig. 1. (a) Face scans are colored according to the values of function f_k at different scales (f_0 being the mean curvature). (b) The 3D frontal acquisition (*subject001* of the UF-3D database) is reported, with the DOG_k values at different scales, and the 3D keypoints detected at that scale (in red). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

value is the maximum with respect to the DOG_k values in the 1-ring neighborhood at the same scale.

Percentage threshold. The extrema of the scale space obtained at the previous step are then sorted according to their magnitude. Only the top 1% of the sorted vertices are retained as extrema in our setting.

Cornerness. The last step, aims to remove unstable extrema, by retaining the features that exhibit corner characteristics. Following [32], this can be done by computing the Hessian at each vertex v of the mesh

$$H(v) = \begin{bmatrix} d_{xx}(v) & d_{xy}(v) \\ d_{yx}(v) & d_{yy}(v) \end{bmatrix}, \tag{2}$$

where d_{xx}, d_{xy}, d_{yx} and d_{yy} are the second partial derivative computed along the x and y directions. In particular, partial derivatives are estimated by applying the definition of directional derivatives given in Eq. (A.1) twice, e.g., $d_{xy} = \nabla_S D_{\vec{x}} f(v) \cdot \vec{y}$, where the gradient is computed using Eq. (A.2). In this context, the directions $\vec{x}$ and $\vec{y}$ represent a local coordinate system in the tangent plane of the vertex v, typically the gradient direction for $\vec{x}$ and its orthogonal direction for $\vec{y}$. The ratio between the largest λ_{max} and the lowest λ_{min} eigenvalues of the Hessian matrix is a good indication of a corner response, which is independent of the local coordinate frame. We typically use $\lambda_{max}/\lambda_{min} = 4$ as a minimum value to threshold responses.

An example of the scale-space construction is reported in Fig. 1. In (a), a sample face scan is colored according to the values of function f_k at different scales (f_0 being the mean curvature). In (b), gray levels are used to represent the DOG values at different scales (i.e., scales 2, 8, 16, 32, 64 and 128 are reported). The Experiment 1 Code Item 2 can Experiment 1 Data Item 1, in order to detect the 3D keypoints and generate the DOG_k images.

2.2. Keypoints distribution

According to an agreed classification [48], meshDOG is an *adaptive-scale* detector, in contrast to *fixed-scale* detectors which find distinctive keypoints at a specific constant scale, given as a parameter to the algorithm. The derivation of multiple DOG scales, allows the identification of more stable keypoints, which are typically located at highest scales, whereas keypoints detected in the first DOG scales are likely to be unstable and more affected by noise. As an example, the keypoints detected at some DOG scales for a sample face scan are highlighted in red in Fig. 1(b). At the first level of the scale-space (see DOG2 in Fig. 1(b)), the keypoints are mainly localized in the mouth and eyes regions (these regions are quite unstable with expressions) and around the nose and the eyebrows (more stable regions under expression changes). As the scale increases, keypoints are extracted by progressively smoothing the mean curvature function, and they tend to be more distributed on the face (see for example DOG64 and DOG128 in Fig. 1(b)). At these latter scales, some keypoints are located in the *forehead*, *cheekbone* and *chin*, with some keypoints close to the *pronasal* and *nasion* (thus, these keypoints are located in regions of the face that are much less affected by expression variations). Some keypoints can be also detected at multiple different scales; in such case, the keypoint occurring at the highest scale is retained. In Fig. 2, two further examples of keypoints detected at different scales are reported.

In general, meshDOG keypoints are located around areas characterized by high local curvature, this being true throughout the different scales. So, their semantic is related to the local curvature properties of the mesh. Our idea is that the robustness of the proposed approach comes from the combination of the presence of many keypoints detected at different scales, with the descriptiveness of local surface features (as discussed in Section 3). The fact that the keypoints are many increases the possibility to have a consistent number of matches also in the case of partial scans. The fact that the keypoints are extracted at different scales increases the probability to have keypoints detected in regions of the face that are not affected by facial expressions so that their descriptors are likely to be not altered in different scans of a same subject. Differently, keypoints detected in noisy regions or regions which are largely affected by expression changes are likely to not match due to their different descriptors. So, our idea is that though individual descriptors are not expression invariant, the overall matching schema can cope with expression variations thanks to the presence of keypoints that are located in regions of the face that are less affected by facial expressions. For the same reason, the approach can cope with missing parts and also occlusions, provided that a sufficient number of matches can be determined between probes and gallery scans. These considerations, motivated us to use the keypoints detected in the last levels of the scale-space. In particular, we considered for the purpose of local descriptor computation only the keypoints that are detected in the last 64 DOG scales (out of the 96 total scales used in the experiments), thus discarding those keypoints that have been detected only in the first 32 scales.

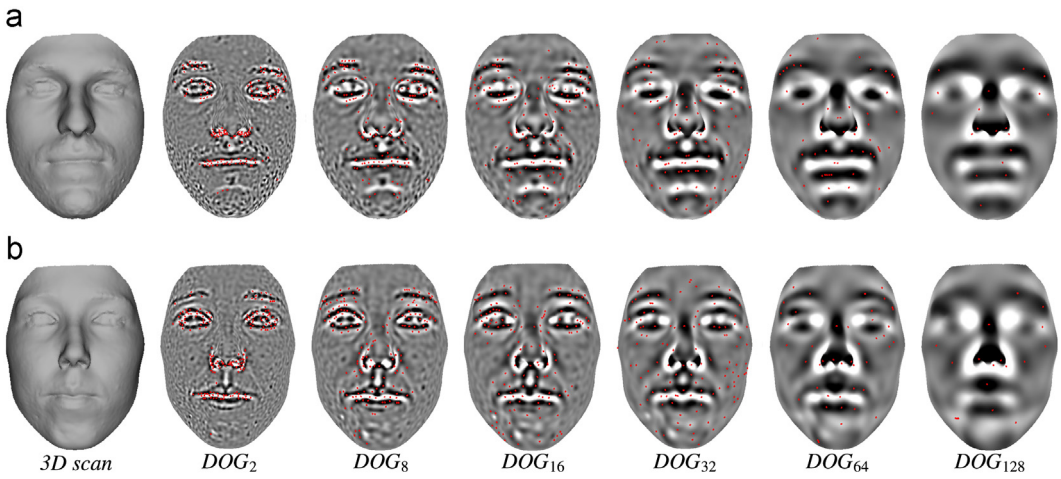


Fig. 2. DOG_k values at different scales and the 3D keypoints detected at that scale for a male subject in (a) and a female subject in (b). (a) subject002 and (b) subject003.

3. Local face descriptors

In order to support face matching, we assume that distinguishing traits of the face can be captured by describing the local morphology of the face in regions centered at 3D keypoints. This approach falls into the category of *signature descriptors* that represent the 3D surface using the neighborhood (called the *support*) of a given keypoint. A common problem faced by these solutions is the need for an invariant local reference frame in order to encode one or more geometric measurements computed individually for each point (vertex) of the support. Typically, the support is a spherical region whose radius determines the level of locality of the descriptor. Small values of the radius yield very local descriptors that capture the shape of the face in small regions around keypoints. By progressively increasing the value of the radius, the descriptor becomes more discriminant, although the probability of including regions of the face affected by undesired artifacts – such as missing parts or deformations caused by facial expressions – increases as well.

Based on these considerations, in the following we propose three different signatures to locally describe the 3D face at the keypoints, namely the *Histogram of Gradients* (HOG) (Section 3.1), the *Histogram of Orientations* (SHOT) (Section 3.2), and the *Geometric Histogram* (GH) (Section 3.3).

3.1. Histogram of gradients

The histogram of gradients descriptor [41] for a vertex extremum v is computed using a support region constituted by the vertices that belong to the neighborhood ring of size r. For each vertex from the neighborhood $v_i \in N_r(v)$, the gradient information $\nabla_S f(v_i)$ is computed using Eq. (A.2). As a first step, a local coordinate system is chosen, in order to make the descriptor invariant to rotation. Then, a histogram of gradient is computed, both spatially, at a coarse level, in order to maintain a certain high-level spatial ordering, and using orientations, at a finer level. Since the gradient vectors are three-dimensional, the histograms are computed in 3D. Since for this descriptor we followed the work of Zaharescu et al. [41], the reader is referred to that work for further implementation details.

3.2. Histogram of orientations

A description of the local shape of the 3D face is accomplished by developing on the idea of the 3D shape context descriptor proposed in [52] and on the work of [53]. The derivation of this signature first requires the definition of a local reference frame capable to make the extracted signature independent from translation and rotation of the mesh.

Local reference frame. In order to guarantee translation and rotation invariance of 3D face description and matching, each local descriptor is computed with respect to a local reference frame determined based on the local morphology of the face. For this purpose, the method proposed in [54] is considered. This avoids the descriptor computation over multiple rotations on different azimuth directions by determining a repeatable normal axis and an unique pair of directions lying on the tangent plane.

Given a keypoint located at vertex v, and a spherical neighborhood of radius R centered on v, a weighted covariance matrix $\mathbf{C}$ of the vertices within the neighborhood is computed as

$$\mathbf{C} = \frac{1}{K} \sum_{i:d_i \le R} (R - d_i)(\mathbf{v}_i - \mathbf{v})(\mathbf{v}_i - \mathbf{v})^T, \tag{3}$$

where $d_i = \|\mathbf{v}_i - \mathbf{v}\|$, and K is a normalization factor computed as

$$K = \sum_{i:d_i \le R} (R - d_i). \tag{4}$$

With respect to the usual computation of the covariance matrix, in Eq. (3) a smaller weight is assigned to distant vertices, and the centroid computation is replaced by the keypoint vertex v. A *total least squares* estimation of the normal direction is obtained by eigenvalue decomposition of the covariance matrix $\mathbf{C}$ of the vertex coordinates within the support. The eigenvectors of $\mathbf{C}$ define repeatable orthogonal directions in the presence of noise and clutter. Eigenvectors of Eq. (3) need to be disambiguated to yield a repeatable local reference frame. The idea is to orient each eigenvector so that its sign is coherent with the majority of the vectors it represents. If the three eigenvectors, given in decreasing eigenvalue order, are indicated as $\mathbf{x}^+$, $\mathbf{y}^+$, and $\mathbf{z}^+$ (and their opposite vectors with $\mathbf{x}^-$, $\mathbf{y}^-$, and $\mathbf{z}^-$), the disambiguated $\mathbf{x}$-axis is defined as

$$\begin{aligned} S_{x^+} &= \{i : d_i \le R \quad \text{and} \quad (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^+ \ge 0\} \\ S_{x^-} &= \{i : d_i \le R \quad \text{and} \quad (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^- > 0\} \\ \mathbf{x} &= \begin{cases} \mathbf{x}^+, & |S_x^+| \ge |S_x^-| \\ \mathbf{x}^-, & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

The same procedure is used to disambiguate the $\mathbf{z}$-axis, whereas the $\mathbf{y}$-axis is obtained as the vector product $\mathbf{z} \times \mathbf{x}$.

Local signature. Once the local reference frame is identified, a spherical support around each keypoint v is considered and the vertices of the mesh included in this spherical region contribute to the computation of the local descriptor. The radial extent of this

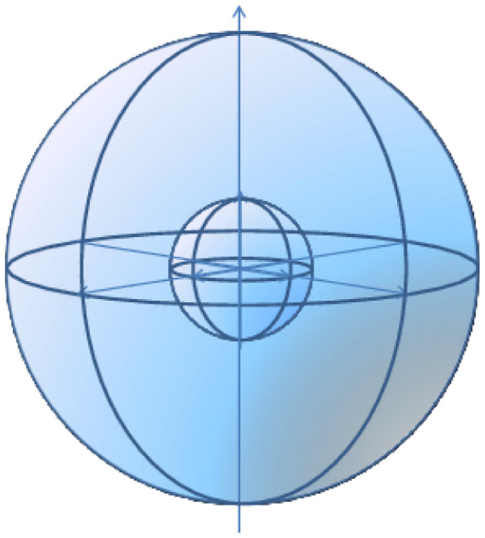


Fig. 3. Spherical local support around a keypoint. The volumetric partition of the sphere along the radial, azimuthal and elevation dimensions is reported.

sphere can be chosen independently from the radius R used for the computation of the local reference frame, but in our solution we considered the spherical support as having the same radius R used for the computation of the reference frame (i.e., 15 mm in our setting). This spherical volume is then divided along three dimensions: *radial*, *azimuthal* and *altitude*.

Along the radial dimension, the sphere is divided into concentric shells. To avoid the quadratic growth of the shell volumes with the shell index, a logarithmic parametrization of the shell radii is used

$$r_i = \frac{1}{s}\log_a\left(a^s\frac{i}{s}\right), \tag{6}$$

where r_i is the radius of the shell of index i, s is the number of shells, and a is a parametrization coefficient that controls the growth of the shell radius (e.g., for $a=1$ the growth is linear, whereas with $a=2$ the volume of the shell is kept constant at different radius). The shells are then divided in the azimuthal plane using sectors of constant angular width, and along the elevation. In the experiments reported in Section 5, we used $a=2$, with three shells, four azimuthal sectors and two divisions along the elevation angle, resulting in a coarse partition of the volume around the keypoint into 24 spatial regions. Fig. 3 shows the idea of the volumetric partitioning of the spherical space around a keypoint (for the clarity of the plot just two shells are reported).

Once the local support is partitioned into volumetric regions (based on the unique 3D local reference frame), the histogram of the normals of the mesh vertices within the support is used as local descriptor (called SHOT in [53]). This histogram based representation provides the filtering effect required to achieve robustness to noise, and enhances the discriminative power of the descriptor by introducing geometric information about the location of the vertices within the support. As final step, all the local histograms are grouped together to form the signature which describes the mesh at the keypoint.

For each of the local histograms, mesh vertices contribute to bins according to a function of the angle θ_i, formed by the normal at each vertex within a volume of the support partition, $\mathbf{n}_{v_i}$, and the normal at the keypoint, $\mathbf{n}_u$. The cos θ_i function is used, in that it can be computed efficiently using the dot product (i.e., $\cos\theta_i = \mathbf{n}_u \cdot \mathbf{n}_{v_i}$), and equally spaced binning on cos θ_i is equivalent to a spatially varying binning on θ_i. This latter property results in a coarser binning for directions close to the reference normal direction and a finer one for orthogonal directions. In this way, small differences in orthogonal directions to the normal that are the most informative ones, cause a vertex to be accumulated in different bins and thus leading to different histograms. Instead, in the presence of quasi-planar regions this choice limits histogram differences due to noise by concentrating the contributions of the vertices in a fewer number of bins. In our experiments, we used 10 bins for each local histogram that combined with the partition into 24 volumetric regions, that results in a 240-dimensional signature for the keypoint.

To avoid boundary effects in the local histograms due to small differences of the spatial subdivision of the support, or to perturbations of the local reference frame, each vertex contributes to four histogram bins according to a quadrilinear interpolation between neighbors bins. In particular, the neighbor bins are represented by the neighboring bin in the local histogram and the bins having the same index in the local histograms of the neighboring volumes of the spatial partition. In doing so, each vertex contributes to neighbors bins by the weight $1-d$, where for the local histogram, d is the distance of the current entry from the central value of the bin; for elevation and azimuth dimensions, d is the angular distance of the entry from the central value of the closer volume along the dimension; for the radial dimension, d is the Euclidean distance of the entry from the central value of the closer volume along the radial dimension. Along each dimension, d is normalized by the distance between two neighbor bins or volumes. Finally, to achieve robustness to variations of the vertex density, all the local histograms are concatenated into a whole descriptor (signature) which is further normalized to sum up to 1, so as to retain the local differences as a source of discriminative information.

The local signature at a generic keypoint is expressed through a normalized histogram $G=(g_1,\ldots,g_N)$ where the size N of the signature depends on the size of the local histograms and on the number of volumes of the partition (i.e., the quantization along the radial, azimuthal and elevation dimension) of the local reference frame ($N=240$ in our case). Given two signatures $G=(g_1,\ldots,g_N)$ and $H=(h_1,\ldots,h_N)$ extracted at two keypoints, their dissimilarity is measured through the *Chi-square* distance χ^2, given by

$$\chi^2(G,H) = \frac{1}{2}\sum_{n=1}^{N}\frac{[g_n-h_n]^2}{g_n+h_n}. \tag{7}$$

The Experiment 2 Code Item 2 can be executed on the Experiment 2 Data Item 1, in order to generate the SHOT signature of a 3D face scan.

3.3. Multi-ring geometric histogram

The geometric histogram (GH) is a local geometric descriptor proposed by Ashbrook et al. [55] and employed in surface alignment and matching. Basically, it is a 2D accumulator, or frequency table that counts the frequencies of two geometrical measurements, namely the angle and the distance between pairs of facets in a given neighborhood of a keypoint. In the following, we propose and describe a variation of the GH, which resulted more suited to our framework. This variant, develops on the idea of constructing the GH descriptor at a given keypoint in an incremental way, by accounting for an ordered sequence of rings defined around the keypoint. This idea is illustrated through the two steps involved in the computation: Derivation of the *ordered ring facets* in the neighborhood of the keypoint; Construction of the *discrete distributions* in each ring. In doing so, it is relevant to note that the GH descriptor is robust to translations and rotations also avoiding the computation of a reference frame.

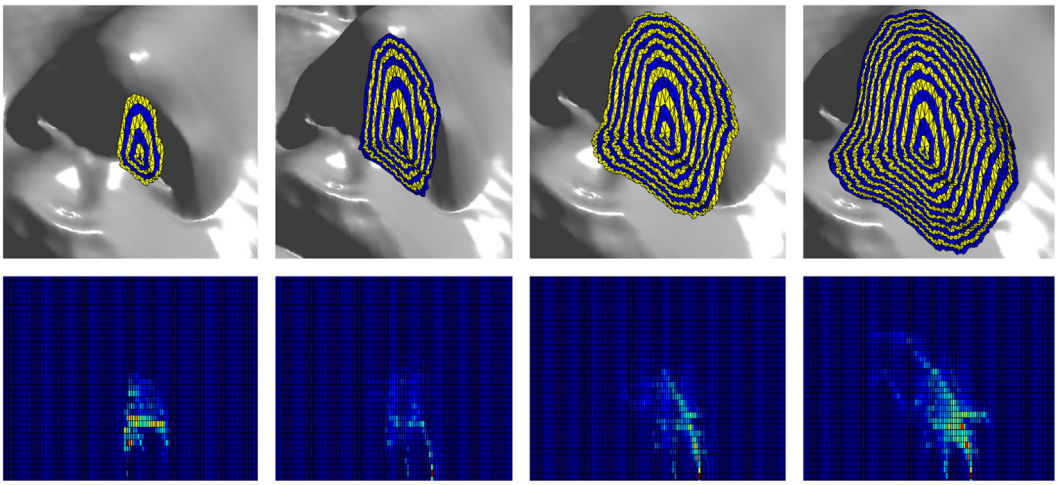
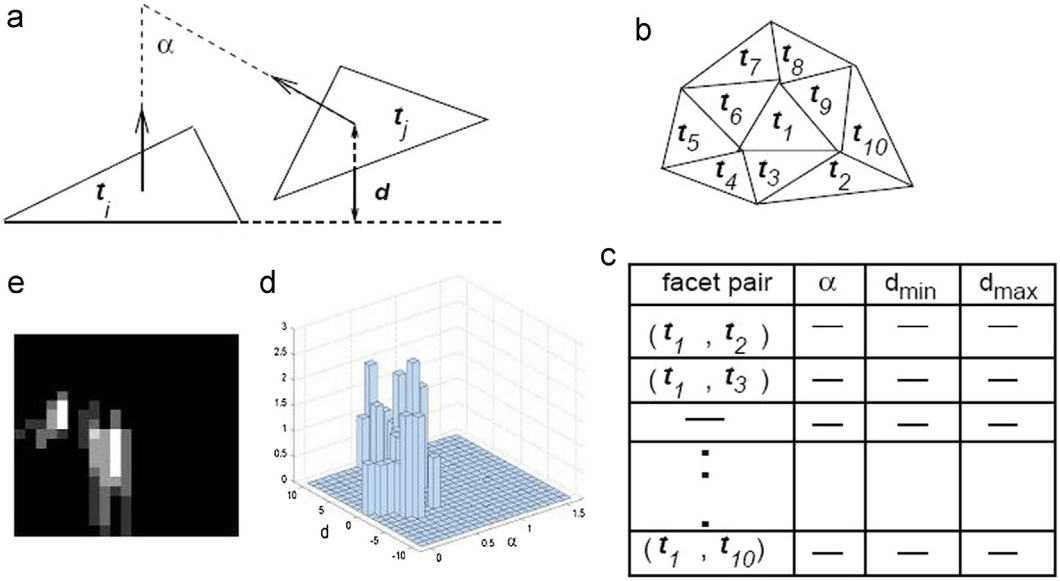


Fig. 4. ORF neighborhoods with different sizes constructed at a facial keypoints near to the nose, and their corresponding GHs.



facet pair	α	d_{min}	d_{max}
(t_1 , t_2)	—	—	—
(t_1 , t_3)	—	—	—
—	—	—	—
⋮			
(t_1 , t_{10})	—	—	—

Fig. 5. (a) The geometric measurements used to characterize the relationship between two facets t_i and t_j. (b) A facet t_1 and its neighbor facets. (c) For each pair (t_1, t_s), $s = 1, \ldots, 10$, 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_s are computed. (d) The pairs (α, d) derived from these measurements are entered in a 2D accumulator, obtaining thus a geometric distribution that characterizes the relationship between the facet t_1 and its neighbors. (e) The geometric distribution can be visualized with a gray level mapping.

Ordered ring facets. The Ordered Ring Facets (ORF) [56] is the method used to identify the facets of the mesh which are comprised in the neighborhood of a keypoint. In this approach, the neighborhood construction around a central facet t_c is performed through a sequence of concentric rings of facets emanating from a root facet (i.e., t_c). The facets are arranged circular-wise within each ring. The size of the neighborhood is simply controlled by the number of rings. This mechanism allows an easy analysis of the GH variability, and thus of the local geometry evolution, as the size of the neighborhood increases. When the triangular mesh is regular and the facets are nearly equilateral, the ORF rings form an approximation of iso-geodesic rings around the central facet t_c. The ORF construction has a linear complexity. Fig. 4 depicts examples of ORF's with increasing number of rings and their related GH's. In the experiments reported in Section 5, we obtained good results by using 8 ORF as neighborhood of the keypoints.

Discrete distribution. Consider a triangular mesh approximation $\hat{S} = \{t_1, \ldots, t_M\}$ of an object surface. The discrete geometric distribution is constructed for each triangular facet t_i in a given mesh which describes its pairwise relationship with each of the other surrounding facets within a predefined neighborhood. The range of the neighborhood controls the degree to which the representation is a local description of shape. Here, we choose a neighborhood range that encompasses the facets that share one or two vertices with the central triangular facet (Fig. 5(b)). The distribution is defined such that it encodes the surrounding shape geometry in a manner which is invariant to rigid transformations of the surface data and which is stable in the presence of surface clutter and missing surface data.

Fig. 5(a) shows the measurements used to characterize the relationship between facet t_i and one of its neighboring facets t_j. These measurements are the relative angle, α, between the facet normals, and the range of perpendicular algebraic distances, d, from the plane in which facet t_i lies to all points on the facet t_j. The range of perpendicular algebraic distances is defined by $[d_{min}, d_{max}]$, where d_{min} and d_{max} are the minimal and the maximal of the distance from the plane, respectively, in which t_i lies to the facet t_j. These extreme entities are simply obtained by calculating the

distances to three vertices of the facet t_j and then selecting the minimal and the maximal distances.

Since the distance measurement is a range rather than a single value, from each measurement $(\alpha, d_{min}, d_{max})$ can be derived a number of measurements (α, d) $(d_{min} \le d \le d_{max})$. This number depends on the amplitude of the range $[d_{min}, d_{max}]$ and the resolution adopted for the distance parameter d. The group of pairs (α, d), extracted from the measurements related to a given facet and its neighbors (Fig. 5(b) and (c)), are entered to a 2D discrete frequency accumulator that encodes the perpendicular distance d and the angle α (Fig. 5(d)). This accumulator has size $N \times M$, where N and M are the number of bins in the axis α and d, respectively. The values of the accumulated matrix are also normalized so as to sum up to 1. The accumulator can be visualized in a 2D plotting using a gray level colormap (Fig. 5(e)), and stored in a matrix for subsequent processing. This representation only depends upon the surface shape and not on the placement of facets over the surface. This independence on the placement of the facets is important as it guarantees the invariance of the correspondence with respect to geometric transformations. A possible variant of the geometric histogram is obtained by considering all the pairs of facets within N_{t_c}, i.e., the set $\{(t_i, t_j),\ t_i \in N_{t_c},\ t_j \in N_{t_c}\}$. The construction of this variant is computationally more demanding as the number of histogram entries evolves quadratically with respect to the number of facets in the neighborhood. Due to this, in our experiment we considered the computation referred to the central facet t_c, using $N=8$ and $M=20$.

With respect to the computation of the central GH, we introduced a variant which is related to the ORF definition. In particular, in our approach, a GH is constructed on each of the rings that constitute the ORF of a keypoint: This means that the GH descriptor is actually given by a set of GH, constructed on the sequence of rings which surround the keypoint. This improves the descriptiveness of GH by capturing information on how the local characteristic of the surface changes when the distance from the keypoint increases. This multi-ring structure is also exploited during the match. In particular, the normalized GH can be viewed as a probability density function, and thus can be adapted to probabilistic matching paradigms. To this end, 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 1D arrays of $K = N \times M$ elements, $A(l) = \{a_1, \ldots, a_K\}$ and $B(l) = \{b_1, \ldots, b_K\}$, their distance at ring-l is computed as

$$d_B(A(l), B(l)) = \sqrt{1 - \sum_{k=1}^{K} \sqrt{(a_k \cdot b_k)}}. \tag{8}$$

The overall distance between two multi-ring GH, computed on L rings is then obtained by accumulating the distances between the GHs at different rings, that is

$$d(A, B) = \sum_{l=1}^{L} d_B(A(l), B(l)). \tag{9}$$

The Experiment 2 Code Item 2 can be executed on the Experiment 2 Data Item 1, in order to generate the GH descriptor of a 3D face scan.

4. Face matching

Given two face scans, their comparison is performed by matching the local shape descriptors at corresponding keypoints under the constraint that a consistent spatial transformation exists between inliers pairs of matching keypoints. To this end, local shape descriptors at the keypoints detected in probe and gallery scans are compared 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 to k_p than to any other keypoint in the probe. For this purpose, distance between keypoints descriptors is measured through the distances presented for the three local descriptor HOG, SHOT and GH, discussed, respectively, in Sections 3.1–3.3. Finally, the candidate matches for which the second best match is significantly worse are accepted (i.e., a match is accepted if the ratio between the distance of the best match and the second best match is lower than 0.7).

This analysis of proximity of keypoint descriptors results in the identification of a candidate set of keypoint correspondences. Identification of the actual set of keypoint correspondences must pass a final constraint targeting the consistent spatial transformation between corresponding keypoints in the probe and gallery scans. The RANSAC algorithm [43,57] is used to identify outliers in the candidate set of keypoint correspondences. 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 transformation (RST transformation). According to this, at each iteration, the RANSAC algorithm validates sampled pairs of matching keypoints under the current RST transformation 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 are removed from the match. Examples of the application of RANSAC are reported in Fig. 6. In the figure, detected keypoints are highlighted with a "+" symbol (in blue); corresponding keypoints based on descriptors matching are connected by green lines; finally, the inliers matching which pass the RANSAC algorithm are shown with a red line connection. It can be observed as by applying the RANSAC algorithm just the matches that show a coherent RST transformation among each other is retained. This avoids matches of keypoints that are located in different parts of the face of two scans. Cases in (a) and (b), respectively, report the match between two scans of the same subject and of different subjects. In Fig. 7, we also report the case in which scans of the same subject with large missing parts (a) and with expression (b) are matched against a full neutral gallery scan. It can be observed as the number of inliers is still high compared to that of different subjects, despite the large missing parts and expression.

Once the set of inlier keypoints is established, the distance between their descriptors is accumulated and averaged. Given a probe and a gallery, the correspondences identified by the spatial transformation hypothesis is a function $\xi : \aleph \mapsto \aleph$ that associates with a keypoint descriptor $\mathcal{C}_k^{(p)}$ in the probe, its corresponding keypoint descriptor $\mathcal{C}_{\xi(k)}^{(g)}$ in the gallery. For each keypoint descriptor in the probe $\mathcal{C}_k^{(p)}$ the distance to the corresponding keypoint descriptor $\mathcal{C}_{\xi(k)}^{(g)}$ in the gallery is evaluated (using Eq. (7) for SHOT or Eq. (9) for GH), and these distances are finally averaged on the total number of inlier matches N_i

$$D = \frac{1}{N_i} \sum_{k=1}^{N_i} \mathcal{D}(\mathcal{C}_k^{(p)}, \mathcal{C}_{\xi(k)}^{(g)}). \tag{10}$$

In this way, the distance between two face scans is regarded as a pair $\langle N_i, D \rangle$. The number of matching inliers is used as measure of distance. In the case two scans have the same number of inliers, the distance D serves as disambiguation value.

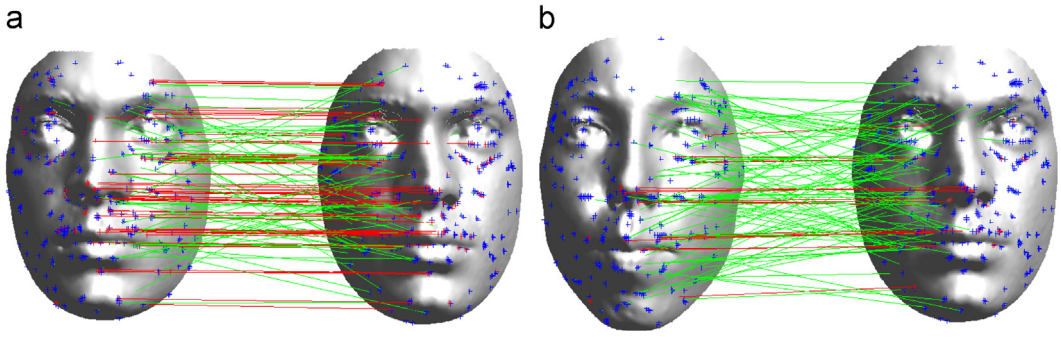


Fig. 6. Matching of scans of same and different subjects are reported in (a) and (b), respectively. All the detected keypoints are shown with "+". Lines indicate matching keypoints (in green), and inliers matching after RANSAC (in red). In the case of scans of the same subject in (a), 61 inlier matches are identified; For scans of different subjects in (b), 18 matches are detected. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

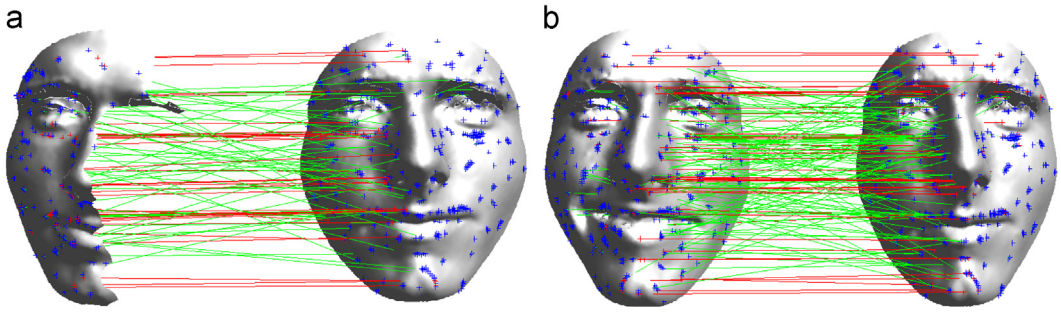


Fig. 7. (a) Partial probe vs. full gallery same subject (34 inliers). (b) Expressive probe vs. neutral gallery same subject (47 inliers). (a) same subject: missing parts and (b) same subject: expression.

The Experiment 3 Code Item 2 can be executed on the Experiment 3 Data Item 1 and Experiment 3 Data Item 2, in order to compute the match between two 3D face scans using local descriptors and RANSAC. An image showing the keypoints matching is also generated.

About the computational complexity of the proposed matching approach, it depends on two main cost factors: the matching of local descriptors and the execution of the RANSAC algorithm. The first term resulted the main source of cost, growing quadratically with the number of keypoints in the two scans. All the three descriptors presented in Section 3 are histogram based and so the complexity in computing their match depends on the distance measure and on the number of histogram bins.

5. Experimental results

The performance of the proposed approach has been evaluated in a comprehensive set of experiments. For the sake of the presentation and discussion, experiments have been divided and organized into two parts:

1. The goal of the first session of experiments was to evaluate the robustness of our 3D face recognition solution to probes showing large facial expressions (from moderate to exaggerated), and extreme pose variations (side rotations of 90°). To this end, experiments were carried out on two datasets that are specifically designed for investigating 3D face recognition in the presence of facial expressions, *The Binghamton University 3D Facial Expression database* (BU-3DFE) [44], and missing parts, *The 2D/3D Florence Face dataset* (UF-3D) [45]. In addition, we provide an in depth investigation on the keypoints detection and repeatability, using the same datasets. Results of this first session of experiments are reported in Section 5.1.
2. In the second session of experiments, the proposed approach is evaluated on a variety of benchmark datasets that differ in the number of scans, acquisition modalities and characteristics of the scans in terms of missing parts, occlusions, and expressions. The used databases are the *Bosphorus* [28], *Gavab* [14] and *UND/FRGC v2.0* [6]. These datasets have been used by many of the existing 3D face recognition works, thus permitting a direct comparison of our approach with state of the art solutions. Section 5.2 reports results of this evaluation.

The datasets listed above largely differ in the scanners used during acquisition (i.e., either laser or structured light scanners), so that both 2.5D (only one z-value is possible at a given xy location) and 3D acquisitions are involved (multiple z-values at the same xy location are allowed). According to this, in the perspective of not to restrict the proposed approach to any particular scenario, in the experimentation we do not make any assumption about the type of scans available in the probe or gallery sets (i.e., they can be either 2.5D or 3D).

5.1. Performance evaluation

The objective of the results reported in this section is to verify the performance of the proposed approach in the case of probes with very large facial expressions (Section 5.1.1), and extreme side rotations (Section 5.1.2). In so doing, we devised an *identification scenario* where the effectiveness of recognition is measured through the rank-k *recognition rate* (RR): a rank-k recognition experiment is successful if the gallery face representing the same individual of the current probe is ranked within the first k positions of the ranked list. The rank-1 value has been reported in our experiments.

5.1.1. The BU-3DFE database

The BU-3DFE database was recently constructed at Binghamton University [44]. It has been designed to provide 3D facial scans of a large population of different subjects each showing a set of facial expressions at various levels of intensity. There are a total of 100 subjects in the database, divided between female (56 subjects) and male (44 subjects). The subjects are well distributed across different ethnic groups or racial ancestries, including *White*, *Black*, *East-Asian*, *Middle-East Asian*, *Hispanic-Latino*, and others. During the acquisition, each subject was asked to perform the neutral facial expression as well as the six basic facial expressions defined by Ekman [58], namely *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. Each facial expression has four levels of intensity, respectively *low*, *middle*, *high* and *highest*, except the neutral facial expression that has only one intensity level. Thus, there are 25 3D facial expression scans for each subject, resulting in 2500 3D facial expression scans in the database. As an example, Fig. 8 shows the 3D scans of a sample subject showing the six basic facial expressions at the *low* and *medium* levels of intensity.

Face recognition results. The BU-3DFE dataset has been used to investigate the robustness of the proposed approach with respect to facial expressions in a wide range of intensity variations, from low to exaggerated. This allowed us to infer some evidence of the facial variations that mostly affect face recognition. So far, the BU-3DFE database has been used mainly to test facial expression recognition methods, rather than the robustness of face recognition methods in the presence of expression variations. Actually, face recognition experiments on the BU-3DFE were conducted in [59,60], but only cumulated results were reported in these works, without a detailed analysis for each expression/intensity. As a consequence, for the large part of the methods reported in the literature, there is no insight of the effect induced by different expressions.

In our experiments, we randomly partitioned the dataset into a training and a testing set. The scans of 20 subjects have been included in the train set and have been used for tuning the parameters of the 3D keypoints detector (i.e., the number of DOG scales, the percentage and cornerness thresholds, see Section 2) and the local descriptors (i.e., number of histogram bins for HOG, SHOT and GH descriptors, see Section 3). A classic grid search approach has been used to this end (this phase is mainly important for keypoints detection, since the percentage and cornerness thresholds largely influence the number of detected keypoints, which can vary of an order of magnitude or so). These parameters have been used in the experiments carried out on this dataset, on the UF-3D database (as reported in the next section) and on the three databases used in Section 5.2. The scans of the remaining 80 subjects have been included in the test set. In particular, we considered the neutral scan of each subject as a reference scan and included it in the gallery set (gallery with 80 neutral scans in total). The probe set is composed of 24 expressive scans for each subject, including for each expression the scans with *low*, *medium high* and *highest* intensity level (see Fig. 8). With this selection, the probe set includes 1920 expressive probe scans. The scans classified as showing a *low* and *medium* expression intensity have moderate and natural expressions, similar to those that are likely to occur in a real context. Instead, scans classified in the BU-3DFE as having *high* and *highest* expression intensity, present quite exaggerated expressions for the large part of the subjects, and are more suited to verify the performance of the approach in very difficult situations.

Using these probe and gallery sets, we performed recognition experiments based on keypoints matching with each of the three local descriptors presented in Section 3. Rank-1 recognition accuracies are reported in Table 1, separately for the six expressions, and for the low and medium intensity level (L1 & L2), and the high and highest level (L3 & L4). From the table, it can be observed that, as the overall performance is concerned, the SHOT descriptor provides the best results among the three local descriptors. Looking in to the performance of the SHOT descriptor, it results that the expression that makes the recognition more difficult is the *surprise* one at L1 & L2. This is confirmed also using the HOG and GH descriptors. This is mainly due to the open mouth

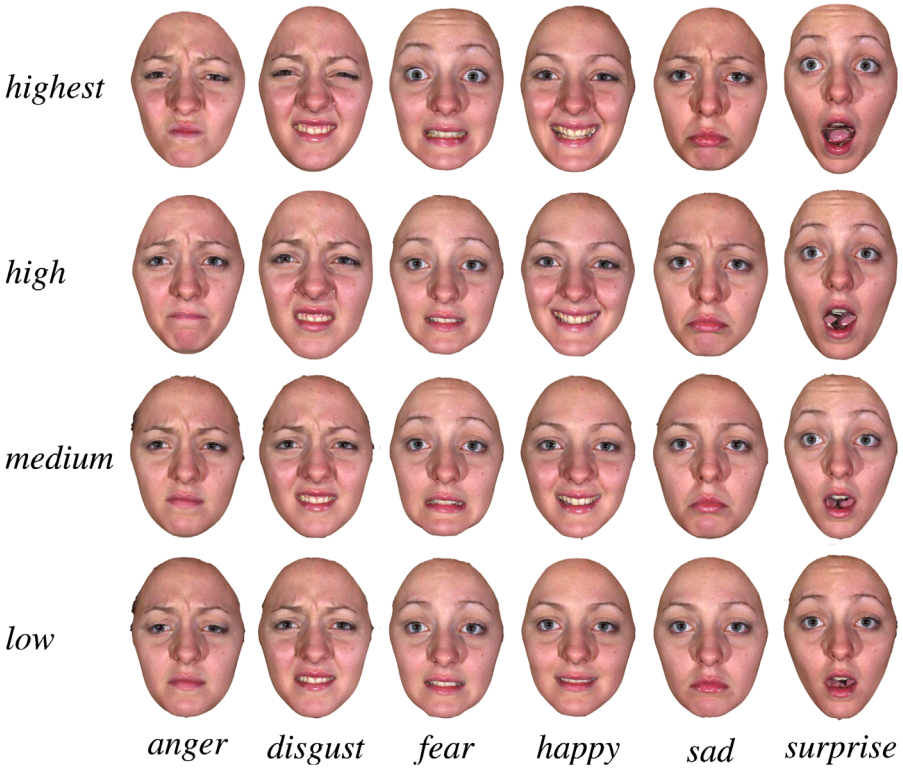


Fig. 8. BU-3DFE: 3D face scans (with texture) of a sample subject showing the six basic facial expressions at the *low*, *medium*, *high* and *highest* level of intensity.

Table 1
BU-3DFE: rank-1 recognition rate (RR) for different expressive scans. Results are reported separately for the HOG, SHOT and GH descriptors. For each descriptor, the average for the *low* and *medium* expression intensity (L1 & L2), and for the *high* and *highest* intensity level (L3 & L4) are reported, together with the average on all the intensity (*All* column).

Expression	Rank-1 RR								
	HOG			**SHOT**			**GH**		
	L1 & L2 (%)	L3 & L4 (%)	All (%)	L1 & L2 (%)	L3 & L4 (%)	All (%)	L1 & L2 (%)	L3 & L4 (%)	All (%)
Angry	90.0	81.3	85.6	93.8	87.5	90.6	90.6	86.3	88.4
Disgust	87.5	75.6	81.6	90.6	78.8	84.7	85.0	79.4	82.2
Fear	88.8	78.8	83.8	91.9	85.6	88.8	84.4	80.0	82.2
Happy	88.1	80.6	84.4	90.0	79.4	84.7	85.6	79.4	82.5
Sad	90.6	82.5	86.6	94.4	90.0	92.2	90.6	85.0	87.8
Surprise	85.0	76.9	80.9	88.8	79.4	84.1	82.5	78.8	80.6
Overall	88.3	79.3	83.8	91.6	83.4	87.5	86.5	81.5	84.0

that appears in the large part of subjects with this expression. The effect of this is a modification of both the location of the detected keypoints with respect to the neutral case, as well as a change of the local descriptors. At L3 & L4 also faces with *disgusted* expression become difficult to be recognized. Furthermore, from this analysis also results that the performance with the GH descriptor seems to degrade more gracefully than for the other descriptors when passing from L1 & L2 to L3 & L4.

5.1.2. The 2D/3D Florence face dataset

The 2D/3D Florence face dataset (UF-2D/3D)[1] has been constructed at the Media Integration and Communication Center of the University of Florence [45]. The dataset consists of high-resolution 3D scans of human faces along with several video sequences of varying resolution and zoom level. This dataset is designed to simulate, in a controlled fashion, realistic surveillance conditions and to test the efficacy of exploiting 3D models in real scenarios. In this work, we used the 3D part of the dataset (UF-3D) that currently includes 53 subjects (14 females and 39 males, numbered from *subject001* to *subject053*) of Caucasian ethnicity. The age of the subjects ranges from 20 to 60, with the majority of the subjects (28) being student at the School of Engineering of the University of Florence, aged between 20 and 30 years. The 3D scans of each subject are acquired in the same session and include two frontal scans with neutral expression (named as *frontal1* and *frontal2*), and two scans where the subject is rotated of 90° on the left and right sides (named *left* and *right*, respectively). In all the acquisitions, the subjects are required to assume a neutral expression, though some scans exhibit moderate, involuntary, facial expressions. The *3dMD face system* [10] scanner has been used in the acquisition, which produces one continuous point cloud from two stereo cameras with a capture speed of about 1.5 ms at the highest resolution, and a geometry accuracy lower than 0.2 mm RMS. As an example, Fig. 9 reports the 3D face scans of two sample subjects.

Face recognition results. The UF-3D dataset allows us to evaluate the recognition accuracy of the proposed solution in the case of frontal neutral probes as well as for probes with extreme yaw rotations. In particular, the left and right probes in this dataset have been acquired with side rotation of 90°, which results in scans with half of the face missing, with consequent very challenging recognition conditions. One neutral scan ("frontal1") has been selected as reference for each subject and included in the gallery. The other neutral scan of each subject ("frontal2") has been used as probe in the "neutral vs. neutral" experiment. The left/right scans have been used in two separate experiments aiming to test the robustness of the proposed approach to partial face matching, where large parts of the face are missing. It is relevant to note that being the proposed approach based completely on 3D processing, both keypoints detection and local description extraction can be performed without the need of costly pose normalization solutions that are required by other existing methods [23,24,29,35].

Results of this evaluation are reported in Table 2. It can be observed that the proposed solution achieves a very high accuracy in matching neutral frontal scans, with each of the three experimented descriptors showing a similar behavior (in this case the SHOT descriptor achieves the best results). For side scans, the accuracy drops significantly with similar results obtained for the left and right scans. The GH descriptor evidences the highest accuracy in this experiment. To the best of our knowledge, the only two other works reporting results on probes with yaw rotations of 90° are those in [36,46], though these two approaches were experimented on the Bosphorus database. Direct comparison of our solution with respect to [36,46] on the Bosphorus database is given in Section 5.2.1.

Fig. 10 shows two examples of wrong recognition for probes with large missing parts. In both the cases, the number of inliers resulted too low to allow rank-1 recognition. For the case on the left, this can be motivated by the presence of a facial expression (see the open mouth) which is combined with a large part of the face missing. In the case on the right, the main problem was originated by the preprocessing operation, which closes holes in the face scans. Due to the large extent of the hole, the hole filling procedure fails in producing a consistent closing, thus altering the face geometry and the keypoints extraction and description.

5.1.3. Localization and repeatability of 3D keypoints

The idea of representing the face by a sparse and adaptive set of automatically detected keypoints relies on the assumption of *intra-subject keypoints repeatability*: Keypoints extracted from different facial scans of the same individual are expected to be located approximately in the same positions of the face. Since keypoints detection only depends on the geometry of the face surface through its mean curvature (see Section 2), these keypoints are not guaranteed to correspond to specific meaningful landmarks of the face. For the same reason, the detection of keypoints on two face scans of the same individual should yield to the identification of the same points of the face, unless the shape of the face is altered by major occlusions or non-neutral facial expressions.

To test the repeatability of keypoints detection, we used the 3D scans of the BU-3DFE database selected for the experiments reported in Section 5.1.1. We followed the approach proposed in [31], and measured the correspondence of the location of keypoints detected in two face scans by performing ICP registration. Accordingly, the 3D faces belonging to the same individual are automatically registered and the errors between the nearest neighbors of their keypoints (one from each face) are recorded. Fig. 11 shows the results of our keypoint repeatability experiment, by reporting the cumulative rate of repeatability as a function of increasing values of the distance. The repeatability reaches a value of 90% for frontal faces with neutral and non-neutral expressions at a distance error of 5 mm (with an average number of 360 keypoints detected per scan). We remark that these results, and those reported in the following about the number of detected keypoints, have been obtained by computing 96 DOG scales, and retaining the unique keypoints that are detected in the last 64 DOG scales (see also Section 2).

[1] The database is publicly available and can be accessed upon request from the following address: http://www.micc.unifi.it/masi/research/ffd/. The dataset is also released within the Elsevier Collage Authoring Environment.

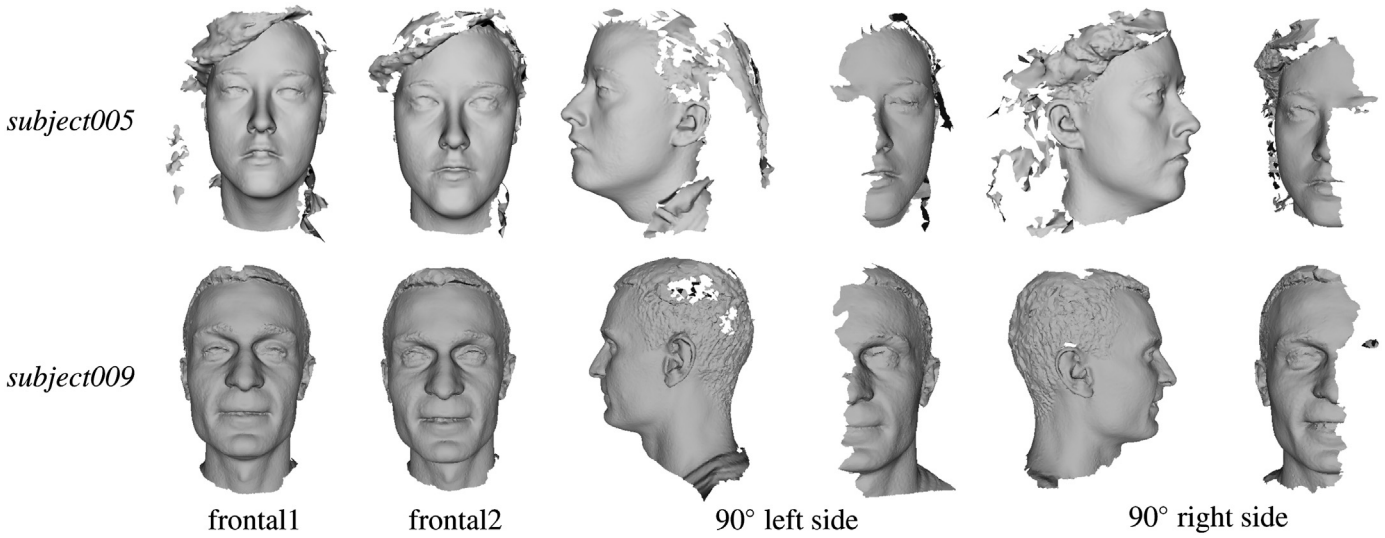


Fig. 9. UF-3D: 3D face scans of two sample subjects. For the left and right cases, the acquired scan is shown as well as its frontal view so has to evidence the missing amount of the facial surface.

Table 2
UF-3D: rank-1 RR for frontal neutral and left/right probes.

Local descriptor	Rank-1 RR	Rank-1 RR		
	Overall (%)	Frontal (%)	Left (%)	Right (%)
HOG	64.8	92.5	49.1	52.8
SHOT	69.2	96.2	54.7	56.6
GH	71.1	94.3	58.5	60.4

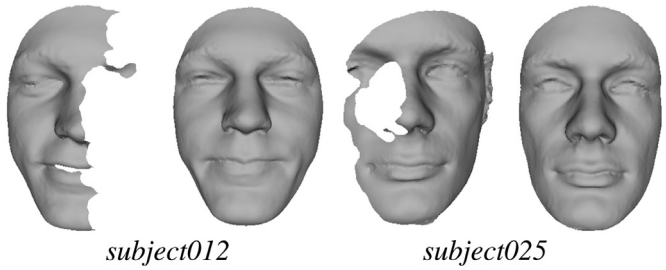


Fig. 10. UF-3D: Example of scans with missing parts that are not recognized when matched against corresponding full gallery scans.

Table 3 also reports the number of keypoints detected on the face scans of the BU-3DFE and the UF-3D datasets. In particular, separate values are given for the *average*, *minimum* and *maximum* number of keypoints. As expected, it can be observed that the largest number of keypoints is detected in the gallery and frontal probes with neutral expression, whereas the number of detected keypoints decreases for side scans. No remarkable differences are observed for the number of keypoints detected on left or right probes. Non-neutral expressions have a small impact on the number of detected keypoints, which remains comparable to that obtained for frontal neutral scans (in some cases, an increase in the number of keypoints is observed).

From Table 3, it results that the number of detected keypoints is quite large. In fact, an important trait of a keypoints detector is the amount of repeatable keypoints it can provide to the subsequent modules of an application. Detecting a small number of keypoints cannot be enough to apply geometrical verification or outliers removal steps, whereas too many may waste computational resources [48]. In the case of meshDOG, the number of detected keypoints is the result of the thresholds involved in the detection algorithm (see Section 2).

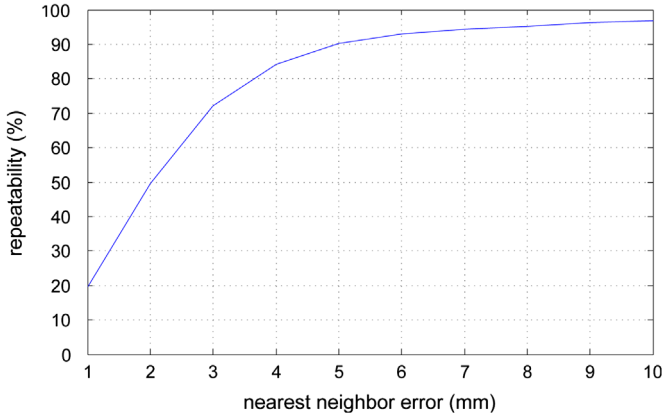


Fig. 11. Repeatability of keypoints.

Table 3
Number of detected keypoints per scan (average, min and max).

Dataset		*Number of keypoints*		
Name	Scans	Avg	Min	Max
UF-3D *frontal*	106	445	346	572
UF-3D *left/right*	106	205	130	396
UF-3D *total*	212	325	130	572
BU-3DFE *neutral*	80	327	265	402
BU-3DFE *expressive*	1920	361	292	464
BU-3DFE *total*	2000	360	265	464

Of course, making these thresholds more selective, the number of keypoints can be reduced. In our experiments, the number of keypoints reported in Table 3 represented a good compromise between computational cost and accuracy of recognition. A number of detected 3D keypoints on 3D face scans of the order of hundreds are also reported for the 3D keypoints detector defined by Mian et al. [31], and for the meshSIFT detector [39,46]. These results seem to support our findings. For example, in the meshSIFT, an average number of about 560 keypoints is reported by the authors, with a number of matching at rank-1 of about 97. The recent survey on the evaluation of 3D keypoint detectors [48], also reported that meshDOG tends to extract a high number of keypoints, that accumulate around areas characterized by high local curvature.

5.2. Comparative evaluation

In this section, the proposed approach is evaluated and compared to state of the art solutions on three benchmark databases: Bosphorus, Gavab and UND/FRGC v2.0. Based on the analysis of Section 5.1, in the following we provide results of our approach only for the GH descriptor. In fact, we found that the GH descriptor provides a good balance of recognition performance between the cases of probes with missing parts and probes with large facial expressions.

5.2.1. The Bosphorus 3D face database

The Bosphorus database has been collected at the Boğaziçi University and made available during 2008 [16]. It consists of 3D facial scans and images of 105 subjects acquired under different expressions and 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 4666 face scans, with the subjects categorized as follows:

- About 34 subjects with up to 31 scans per subject (including 10 expressions, 13 poses, 4 occlusions and 4 neutral).
- About 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, which provide more realistic and pronounced expressions.

Face recognition results and comparative evaluation. In our experiments, we used the same experimental protocol proposed in [36,46], 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 classes as reported in Table 4 (the number of probes per class is also indicated). The first class groups probe according to their facial expression, distinguishing between neutral probes and expressive probes categorized according to the six expressions defined by Ekman [58], plus some not-classified probes. Probes where subjects exhibit face action units are accounted in the second class, by considering scans with Lower Face Action Unit (LFAU), Upper Face Action Unit (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 the methods in [36,46] we provide the rank-1 RR accuracy as reported in the respective publications.

Table 4
Bosphorus: rank-1 RR for different probe classes. Results of our approach are compared with those reported in [36,46].

Probes (#)	Li et al. [36] % rank-1RR	Smeets et al. [46] % rank-1RR	**This work** rank-1RR
Neutral (194)	100.0	–	97.9
Anger (71)	88.7	–	85.9
Disgust (69)	76.8	–	81.2
Fear (70)	92.9	–	90.0
Happy (106)	95.3	–	92.5
Sad (66)	95.5	–	93.9
Surprise (71)	98.6	–	91.5
Other (18)	–	–	100.0
LFAU (1549)	97.2	–	96.5
UFAU (432)	99.1	–	98.4
CAU (169)	98.8	–	95.6
YR (735)	78.0	–	81.6
PR (419)	98.8	–	98.3
CR (211)	94.3	–	93.4
O (381)	99.2	–	93.2
All (4561)	94.1	93.7	93.4

From the table, we first note that the approach by Li et al. [36] reports a detailed analysis for the different probe categories, whereas in Smeets et al. [46] results are presented in a cumulative way. Results show that our approach has overall performance which are very close to state of the art solutions, and for some category are even better. In particular, our solution performs particularly well in recognizing scans with missing parts (see for example the YR category). More in detail, our approach achieves an accuracy of 45.7% on scans with $\pm 90^\circ$ left/right yaw rotations. Results for these scans are not reported directly in [46]. However, authors also reported the overall recognition in the case the $\pm 90^\circ$ scans are removed. So, it is possible to derive the accuracy of [46] on $\pm 90^\circ$ scans to be around 25%.

We guess the lower performance achieved in [46] on scans with very large missing parts are mainly due to the way local descriptors are computed. In fact, in [46] the local support used for the computation of the meshSIFT feature is quite large and increases with the scale at which keypoints are detected. As a result, keypoints detected at the highest scales, which in principle are the most stable, have local descriptors which span a large part of the face. This reduces the robustness of the descriptor to missing parts. In our case instead, the local support is quite small thanks to the descriptive capability of the multi-ring GH descriptor, thus making our representation quite robust to missing parts of the face.

5.2.2. Gavab database

The Gavab database [14] comprises facial scans with large pose and expression variations, and noisy acquisitions. It includes 3D face scans of 61 adult Caucasian individuals (45 males and 16 females). For each individual, nine scans are taken that differ in the acquisition viewpoint and facial expressions, resulting in a total of 549 facial scans. In particular, for each individual, there are two frontal face scans with neutral expression, two face scans where the subject is acquired with a rotated posture of the face (around $\pm 35^\circ$ looking-up or looking-down) and neutral facial expression, and three frontal scans in which the person laughs, smiles, or shows a random expression. Finally, there are also two side scans nominally acquired with a rotation of $\pm 90^\circ$ left and right. In our experiments, we used all the probes and compared them against the gallery scans. The gallery includes, for each subject, the scan named "frontal1" according to the experimental protocol of this dataset.

Face recognition results and comparative evaluation. On this dataset, our results are compared with those reported in [29,35] that used a similar experimental setup. Table 5 summarizes the evaluation using rank-1 RR. Results demonstrate that our approach is capable of achieving or improving state of the art performance for all the classes of scans. As a general behavior, a quite large difference in recognizing left and right side scans can be noted for this dataset (about 10%, 14% and 16% decrease, respectively, for our work and the approaches in [29,35]). Measuring the yaw rotation for the left and right side scans, we obtained an average angle of about 50° and 70°, respectively. These rotation angles are lower than the nominal values reported in the database description, and the difference of around 20° between left and right rotations motivate the different recognition accuracy in the two cases.

Table 5
Gavab dataset: Comparison between methods reporting partial face matching results on left/right scans. The rank-1 RR is reported (highest RR values are evidenced in bold for each class).

Dataset		Rank-1 RR		
Name	Scans	Drira et al. [29] (%)	Huang et al. [35] (%)	**This work (%)**
Frontal neutral	61	**100.0**	**100.0**	**100.0**
Frontal expressive	183	**94.5**	94.0	94.0
Neutral + expressive	244	94.7	**95.5**	95.1
Looking-down	61	**100.0**	96.7	95.1
Looking-up	61	**98.4**	96.7	96.7
Left side	61	86.9	**93.4**	**93.4**
Right side	61	70.5	78.7	**83.6**

5.2.3. UND/FRGC v2.0 database

We performed experiments on the side facial scans of the ear database from the *University of Notre Dame* (UND) [15], collections F and G. This database was created for ear recognition purposes and contains side scans with yaw rotations of 45°, 60° and 90°. Similarly to [23], we used the 45° side scans (119 subjects, with 119 left and 119 right scans) and the 60° side scans (88 subjects, with 88 left and 88 rights scans). As noted in [23], even if these side scans are marked as 45° and 60° by the creators of the database, the measured average yaw angle of rotation is 65° and 80°, respectively. There is a partial overlap between subjects in the UND and in the FRGC v2.0 databases, but not all subjects exist in both the UND and FRGC v2.0. In fact, the number of common subjects between the gallery scans (i.e., frontal scans in the FRGC v2.0) and the 45° side scans is 39, and between the gallery scans and the 60° side scans is 33. According to the partition of the probes used in [23], in our experiments we considered the following test datasets:

- DB45F: Gallery set has one frontal scan for each of the 466 subjects of the FRGC v2.0; Probe set has 45° left/right side scans for each of the 39 subjects.
- DB60F: Gallery set has one frontal scan for each of the 466 subjects of the FRGC v2.0; Probe set has 60° left/right side scans for each of the 33 subjects.

In both the cases, there is only one gallery scan per subject (466 scans in total), and the gallery coincides with that of the FRGC v2.0 dataset. In addition, all the subjects included in the probe set are also present in the gallery set (the opposite is not always true). In the following, we will also use UND45 left/right and UND60 left/right to refer to the probe sets constituted by the 45° left/right side scans and by the 60° left/right side scans, respectively.

Face recognition results and comparative evaluation. In the following, we compare the proposed solution with the approaches in [23] (*automatic* and *manual*) and [24] that have been evaluated on the UND/FRGC v2.0 following the same experimental setup and protocol. Results of the comparative evaluation are summarized in Table 6 using rank-1 RR. Results are organized in three parts:

- UND45 left/right: At rank-1 the approach in [23] (*manual*) results the most effective. We point out that the solution in [23] can use both automatically and manually detected facial landmarks in order to identify face regions used for face alignment and recognition. Quite interestingly, the accuracy of our solution is very close to the accuracy of the solution relying on manual annotation [23], and higher than the accuracy of the solution relying on automatic detection.

Table 6
UND dataset: Comparison between methods reporting partial face match results on the left and right scans of the UND probes. The RR at rank-1 is reported, with values for individual experiments and their average (*avg*). The highest RR values for each dataset are reported in bold.

Dataset		Rank-1 RR			
		Perakis et al. [23]		Passalis et al. [24] (%)	**This work (%)**
Name	Scans	*Manual* (%)	*Automatic* (%)		
UND45 *left*	39	**92.3**	74.4	–	87.2
UND45 *right*	39	**82.1**	64.1	–	**82.1**
UND45 *avg*	78	**87.2**	69.2	–	84.6
UND60 *left*	33	42.4	42.4	–	**66.7**
UND60 *right*	33	42.4	45.5	–	**69.7**
UND60 *avg*	66	42.4	43.9	–	**68.2**
UND *left avg*	72	69.4	59.7	74.6	**77.8**
UND *right avg*	72	63.9	55.6	**78.9**	76.4
UND *total avg*	144	66.7	57.6	76.8	**77.1**

- UND60 left/right: These results evidence the large improvement in the recognition accuracy (more than 20% at rank-1) that our approach achieves with respect to the other solutions.
- UND left/right (45° plus 60°), UND total: Overall, at rank-1, our approach is competitive with the state of the art solution recently reported in [24].

The comparative evaluation evidences that our solution is capable of achieving and in some cases improve state of the art results in the recognition of partial face scans. This is obtained with a completely automatic solution and at a reasonable computational cost. We also evidence that, unlike the solution in [24], our approach does not rely on any assumption of symmetry of the face to reconstruct its global geometry, but only relies on the match of descriptors extracted at detected keypoints of existing parts of the face. This makes our solution more generally applicable.

6. Discussion and conclusions

In this work, we have proposed an original approach to 3D face recognition based on the idea of capturing local information of the face surface around a set of 3D keypoints detected at multiple scales according to differential surface measurements. The approach, first detects 3D keypoints of the face mesh, then local descriptors are extracted at each keypoint and used to find keypoint correspondences during the match. The approach makes no assumption about the correspondence of detected keypoints to specific landmarks on the face, and therefore it can support the comparison of probe and gallery scans even in the case probe scans represent just a part of the face. To improve the accuracy of keypoints correspondences, a spatial constraint is introduced using the RANSAC algorithm.

A preliminary evaluation carried out on the BU-3DFE and the UF-3D datasets showed the viability of the approach in managing moderate as well as exaggerated facial expressions and extreme rotations of the scans, with consequent absence of large parts of the face. This first round of experiments suggested us to use the multi-ring GH descriptor in the subsequent comparative evaluation that has been extended to the Bosphorus, Gavab and UND/FRGC v2.0 databases. Results of this comparison showed that our solution can compete with state of the art works evidencing a

clear advantage in the case of probes with large missing parts. In summary, our view is that the proposed approach presents some interesting solutions in the perspective to make 3D face recognition deployable in real non-cooperative context of use: The approach is fully-3D, reducing to the minimum the need for preprocessing operations, not requiring any costly normalization or alignment; The meshDOG keypoints combined with the multi-ring GH descriptor as proposed in this work, provide a good compromise between robustness to expression changes and missing parts of the face; The inclusion of a statistical technique for outlier removal of matching keypoints largely improves the recognition results.

In perspective, the proposed approach could be further improved by fusing together the local descriptors proposed in this work so as to exploit and combine their strengths. Furthermore, the proposed framework can be easily adapted to include texture information of the face surface, so as to define a multi-modal solution that can combine together in a *native* way (i.e., at the level of the function used for meshDOG detection) 2D and 3D data.

Acknowledgments

The authors thank Iacopo Masi at the University of Firenze for making available the 2D/3D Florence face database, and Emiliano Mazzoncini at the University of Firenze for writing part of the code for meshDOG/meshHOG keypoint detection and description.

Appendix A. Operations on the mesh

In order to make this work self-comprehensive, in the following we summarize the main operations performed on the mesh surface that we used in the paper (according to the analysis in [41]). In so doing, we consider uniformly sampled triangulated meshes S, that is meshes whose facets are triangles of approximately the same area and whose vertices have a valence close to 6 (the vertex's valence being defined as the number of edges incident on it). Simple mesh operations can be applied to transform a non-uniform mesh into a uniform one [61].

A mesh S is viewed as a pair $\langle V, E\rangle$, where $V=\{v_i\}_{i=1,\ldots,N}$ is the set of mesh vertices (with $\mathbf{v_i}$ we indicate the 3D point associated to the vertex v_i, i.e., $\mathbf{v}_i \in \mathcal{R}^3$), and $E=\{e_{ij}\}$ is the set of mesh edges between adjacent vertices. The ring of a vertex $ring(v_i, n)$ is the set of vertices that are at distance n from v_i on S, where the distance n is the minimum number of edges between two vertices. Thus $ring(v_i, 0)$ is the vertex v_i itself, and $ring(v_i, 1)$ is the set of direct neighbors of v_i. According to this, the neighborhood $N_n(v_i)$ is the set of rings $\{ring(v_i, k)\}_{k=0,\ldots,n}$. We further denote $\vec{\mathbf{n}}_{v_i}$ the unit vector normal to the surface S at vertex v_i, computed as the average direction of the normals of the triangles incident to v_i.

Given a scalar function f defined on the vertices of a mesh S, that is $f: S \to \mathcal{R}$, the operations of *directional derivative*, *gradient* and *convolution* of f on the *discrete domain* of the vertices of S can be computed as reported in the following.

Discrete directional derivative. The discrete directional derivative of f on S along the direction of the edge $\vec{e_{ij}}$ (i.e., the direction of the vector $\overrightarrow{v_i v_j}$ originating in v_i and oriented from v_i to v_j) is defined as

$$D_{\vec{e_{ij}}} f(\mathbf{v}_i) = \frac{1}{\|\mathbf{v}_i - \mathbf{v}_j\|} \cdot (f(\mathbf{v}_j) - f(\mathbf{v}_i)), \tag{A.1}$$

with $v_j \in ring(v_i, 1)$, and using the fact that up to the first order $f(\mathbf{v}_j) - f(\mathbf{v}_i) = \nabla_S f(\mathbf{v}_i) \cdot (\mathbf{v}_j - \mathbf{v}_i)$ around v_i.

Discrete gradient. The gradient operator $\nabla_S f(\mathbf{v}_i)$ of f at vertex $\mathbf{v}_i \in S$ is defined as (based on the directional derivatives on v_i)

$$\nabla_S f(\mathbf{v}_i) = \sum_{v_j \in ring(v_i,1)} (w_{ij} \cdot D_{\vec{e_{ij}}} f(\mathbf{v}_i)) \cdot \overrightarrow{u_{ij}}, \tag{A.2}$$

where w_{ij} weights the contribution of $D_{\vec{e_{ij}}}$ and $\overrightarrow{u_{ij}}$ is the normalized projected direction of $\overrightarrow{v_{ij}}$ in the tangent plane at v_i. Assuming that S is uniformly sampled and thus that neighbors around v_i are equally spaced we get: $w_{ij} = 1/val(v_i)$ where $val(v_i)$ is the valence of v_i (i.e., the number of edges incident on it). For non-uniformly sampled meshes, the weights are a function of the angles between the directions $\overrightarrow{u_{ij}}$ around v_i in the tangent plane at v_i.

Discrete convolution. The convolution of the function f with a kernel h on S is defined as

$$(f * h)(v_i) = \frac{1}{H} \cdot \sum_{v_j \in N_n(v_i)} h(\|\mathbf{v}_i - \mathbf{v}_j\|) \cdot f(\mathbf{v}_j), \tag{A.3}$$

where the kernel weighs the neighboring vertices v_j as a function of their distances from vertex v_i, and $H = \sum_{v_j \in N_n(v_i)} h(\|\mathbf{v}_i - \mathbf{v}_j\|)$ is a normalization factor. Notice that, as for the discrete gradient, a uniformly sampled mesh is assumed. As a consequence, contributions of neighboring vertices v_j in the above expression are equally weighted with respect to their spatial arrangements. In this work, we used the above definition with the first ring only (i.e., $n=1$, so that the vertex v_i and the vertices in its $ring(v_i, 1)$ are considered).

Appendix. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.cag.2013.04.001.

Note from publisher: this material was originally submitted as part of the Collage Executable Paper pilot, please visit http://www.elsevier.com/executablepaper for more information.

References

[1] Phillips PJ, Scruggs WT, O'Toole AJ, Flynn PJ, Bowyer KW, Schott CL, Sharpe M. FRVT 2006 and ICE 2006 large-scale results. In: Technical Report, NISTIR 7408, National Institute of Standards and Technology; 2007.
[2] Bowyer KW, Chang KI, Flynn PJ. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. Comput Vision Image Understanding 2006;101(1):1–15.
[3] Berretti S, Del Bimbo A, Pala P. 3D face recognition using iso-geodesic stripes. IEEE Trans Pattern Anal Mach Intell 2010;32(12):2162–77.
[4] Mian AS, Bennamoun M, Owens R. An efficient multimodal 2D-3D hybrid approach to automatic face recognition. IEEE Trans Pattern Anal Mach Intell 2007;29(11):1927–43.
[5] Wang Y, Liu J, Tang X. Robust 3D face recognition by local shape difference boosting. IEEE Trans Pattern Anal Mach Intell 2010;32(12):1858–70.
[6] Phillips PJ, Flynn PJ, Scruggs T, Bowyer KW, Chang J, Hoffman K, Marques J, Min J, Worek W. Overview of the face recognition grand challenge. In: Proceedings of the IEEE workshop on face recognition grand challenge experiments. San Diego, CA; 2005. p. 947–54.
[7] Berretti S, Del Bimbo A, Pala P. Distinguishing facial features for ethnicity-based 3D face recognition. ACM Trans Intell Syst Technol 2012;3(3):1–20.
[8] Ballihi L, Ben Amor B, Daoudi M, Srivastava A, Aboutajdine D. Boosting 3D-geometric features for efficient face recognition and gender classification. IEEE Trans Inf Forensics Secur 2012;7(6):1766–79.
[9] Artec. ⟨http://www.artec3d.com⟩.
[10] 3dMD. ⟨http://www.3dmd.com⟩.
[11] Kinect. ⟨http://www.xbox.com⟩.
[12] Berretti S, Del Bimbo A, Pala P. Superfaces: a super-resolution model for 3D faces. In: Proceedings of the workshop on non-rigid shape analysis and deformable image alignment (NORDIA'12). Firenze, Italy; 2012. p. 73–82.
[13] Sandbach G, Zafeiriou S, Pantic M, Rueckert D. Recognition of 3D facial expression dynamics. Image Vision Comput 2012;30(10):762–73.
[14] Moreno AB, Sánchez Á. Gavabdb: A 3D face database. In: Proceedings of the workshop on biometrics on the Internet. Vigo, Spain; 2004. p. 75–80.
[15] University of Notre Dame, University of Notre Dame biometrics database; 2008. ⟨http://www.nd.edu/@cvrl/UNDBiometricsDatabase.html⟩.
[16] Savran A, Alyüz N, Dibeklioğlu H, Çeliktutan O, Gökberk B, Sankur B, Akarun L. Bosphorus database for 3D face analysis. In: Proceedings of the first COST 2101 workshop on biometrics and identity management; 2008.

[17] Besl PJ, Mc Kay ND. A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Intell 1992;14(2):239–56.
[18] Lu X, Jain AK, Colbry D. Matching 2.5D face scans to 3D models. IEEE Trans Pattern Anal Mach Intell 2006;28(1):31–43.
[19] Ben Amor B, Ardabilian M, Chen L. New experiments on ICP-based 3D face recognition and authentication. In: Proceedings of the international conference on pattern recognition (ICPR'06), vol. 3. Hong Kong; 2006. p. 1195–9.
[20] Bronstein AM, Bronstein MM, Kimmel R. Robust expression-invariant face recognition from partially missing data. In: Proceedings of the European conference on computer vision. Gratz, Austria; 2006. p. 396–408.
[21] Wang Y, Tang X, Liu J, Pan G, Xiao R. 3D face recognition by local shape difference boosting. In: Proceedings of the European conference on computer vision, vol. 1. Marseille, France; 2008. p. 603–16.
[22] Colombo A, Cusano C, Schettini R. Gappy pca classification for occlusion tolerant 3D face detection. J Math Imaging Vision 2009;35(3):193–207.
[23] Perakis P, Passalis G, Theoharis T, Toderici G, Kakadiaris IA. Partial matching of interpose 3D facial data for face recognition. In: Proceedings of the international conference on biometrics: theory, applications, and systems. Washington, DC; 2009. p. 1–8.
[24] Passalis G, Perakis P, Theoharis T, Kakadiaris IA. Using facial symmetry to handle pose variations in real-world 3D face recognition. IEEE Trans Pattern Anal Mach Intell 2011;33(10):1938–51.
[25] Kakadiaris IA, 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 2007;29(4):640–9.
[26] Chang KI, Bowyer KW, Flynn PJ. Multiple nose region matching for 3D face recognition under varying facial expression. IEEE Trans Pattern Anal Mach Intell 2006;28(6):1695–700.
[27] Faltemier TC, Bowyer KW, Flynn PJ. A region ensemble for 3D face recognition. IEEE Trans Inf Forensics Secur 2008;3(1):62–73.
[28] Alyüz N, Gökberk B, Akarun L. 3D face recognition system for expression and occlusion invariance. In: Proceedings of the IEEE international conference on biometrics: theory, applications, and systems. Washington, DC, USA; 2008. p. 1–7.
[29] Drira H, Ben Amor B, Daoudi M. A. Srivastava, Pose and expression-invariant 3D face recognition using elastic radial curves. In: Proceedings of the British machine vision conference. Aberystwyth, UK; 2010. p. 1–11.
[30] Gupta S, Markey MK, Bovik AC. Anthropometric 3D face recognition. Int J Comput Vision 2010;90(3):331–49.
[31] Mian AS, Bennamoun M, Owens R. Keypoint detection and local feature matching for textured 3D face recognition. Int J Comput Vision 2008;79(1):1–12.
[32] Lowe D. Distinctive image features from scale-invariant key points. Int J Comput Vision 2004;60(2):91–110.
[33] Mian AS, Bennamoun M, Owens R. On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes. Int J Comput Vision 2010;89(2–3):348–61.
[34] Huang D, Zhang G, Ardabilian M, Wang Y, Chen L. 3D Face Recognition using distinctiveness enhanced facial representations and local feature hybrid matching. In: Proceedings of the IEEE international conference on biometrics: theory, applications and systems (BTAS'10). Washington DC, USA; 2010. p. 1–7.
[35] 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 2012;7(5):1551–64.
[36] 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 the IEEE international conference on image processing (ICIP'11); 2011. p. 3053–6.
[37] Claes P, Smeets D, Hermans J, Vandermeulen D, Suetens P. SHREC'11 track: robust fitting of statistical model. In: Proceedings of the eurographics workshop on 3D object retrieval. Llandudno, UK,;2011. p. 89–95.
[38] Veltkamp R, van Jole S, Drira H, Ben Amor B, Daoudi M, Li H, Chen L, Claes P, Smeets D, Hermans J, Vandermeulen D, Suetens P. SHREC'11 track: 3D face models retrieval. In: Proceedings of the eurographics workshop on 3D object retrieval. Llandudno, UK; 2011. p. 89–95.
[39] Maes C, Fabry T, Keustermans J, Smeets D, Suetens P, Vandermeulen D. Feature detection on 3D face surfaces for pose normalisation and recognition. In: Fourth IEEE international conference on biometrics: theory applications and systems (BTAS); 2010. p. 1–6.
[40] Tola E, Lepetit V, Fua P. A fast local descriptor for dense matching. In: Proceedings of the international conference on computer vision and pattern recognition. Anchorage, AK; 2008. p. 1–8.
[41] Zaharescu A, Boyer E, Varanasi K, Horaud R. Surface feature detection and description with applications to mesh matching. In: Proceedings of the IEEE international conference on computer vision and pattern recognition. Miami Beach, FL; 2009. p. 373–80.
[42] Zaharescu A, Boyer E, Horaud R. Keypoints and local descriptors of scalar functions on 2D manifolds. Int J Comput Vision 2012;100(1):78–98.
[43] Zuliani M, Kenney CS, Manjunath BS. The multiransac algorithm and its application to detect planar homographies. In: Proceedings of the IEEE international conference on image processing; 2005. p. 153–6.
[44] Yin L, Wei X, Sun Y, Wang J, Rosato M. A 3D facial expression database for facial behavior research. In: Proceedings of the IEEE international conference on automatic face and gesture recognition. Southampton, UK; 2006. p. 211–6.
[45] Bagdanov AD, Del Bimbo A, Masi I. The Florence 2D/3D hybrid face dataset. In: Proceedings of the joint ACM workshop on human gesture and behavior understanding (J-HGBU'11). Arizona, USA; 2011. p. 79–80.
[46] Smeets D, Keustermans J, Vandermeulen D, Suetens P. meshSIFT: Local surface features for 3D face recognition under expression variations and partial data. Comput Vision Image Understanding 2013;117(2):158–69.
[47] Boyer E, Bronstein AM, Bronstein MM, 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 the eurographics workshop on 3D object retrieval (3DOR 2011). Llandudno, UK; 2011.
[48] Salti S, Tombari F, Di Stefano L. Performance evaluation of 3D keypoint detectors. Int J Comput Vision 2013;102(2–3):198–220.
[49] Sipiran I, Bustos B. A robust 3D interest points detector based on Harris operator. In: Proceedings of the eurographics workshop on 3D object retrieval, eurographics association. Norrköping, Sweden; 2010. p. 7–14.
[50] Litman R, Bronstein AM, Bronstein MM. Diffusion-geometric maximally stable component detection in deformable shapes. Comput Graph 2011;35(3):549–60.
[51] Peyre G. Toolbox graph. In: MATLAB central file exchange select; 2009.
[52] Frome A, Huber D, Kolluri R, Bülow T, Malik J. Recognizing objects in range data using regional point descriptors. In: Proceedings of the European conference on computer vision, vol. 3. Prague, Czech Republic; 2004. p. 224–37.
[53] Tombari F, Salti S, Di Stefano L. Unique signature of histograms for local surface description. In: European conference on computer vision, vol. III. Heraklion, Crete, Greece; 2010. p. 347–60.
[54] Tombari F, Salti S, Di Stefano L. Unique shape context for 3D data description. In: Proceedings of the ACM workshop on 3D object retrieval. Firenze, Italy; 2010. p. 57–62.
[55] Ashbrook A, Fisher R, Robertson C, Werghi N. Finding surface correspondence for object recognition and registration using pairwise geometric histograms. In: Proceedings of the European conference on computer vision. Friburg, Germany; 1998. p. 674–86.
[56] 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;2012(144):1–20.
[57] Fischler MA, Bolles RC. Random sample consensus. Commun ACM 1981;24(6):381–95.
[58] Ekman P. Universals and cultural differences in facial expressions of emotion. In: Proceedings of the of the Nebraska symposium on motivation, vol. 19. Lincoln, NE; 1972. p. 207–83.
[59] Mpiperis I, Malassiotis S, Strintzis M. Bilinear models for 3-D face and facial expression recognition. IEEE Trans Inf Forensics Secur 2008;3(3):498–511.
[60] Ocegueda O, Passalis G, Theoharis T, Shah S, Kakadiaris I. UR3D-C: linear dimensionality reduction for efficient 3d face recognition. In: International joint conference on biometrics (IJCB'11). Washington DC, USA; 2011. p. 1–6.
[61] Kobbelt L, Bareuther T, Seidel H-P. Multiresolution shape deformations for meshes with dynamic vertex connectivity. Eurographics 2000;19(3):1–11.