Int J Comput Vis
DOI 10.1007/s11263-014-0785-6

Towards 3D Face Recognition in the Real: A Registration-Free Approach Using Fine-Grained Matching of 3D Keypoint Descriptors

Huibin Li · Di Huang · Jean-Marie Morvan · Yunhong Wang · Liming Chen

Received: 26 April 2013 / Accepted: 27 October 2014

Abstract Registration algorithms performed on point clouds or range images of face scans have been successfully used for automatic 3D face recognition under expression variations, but have rarely been investigated to solve pose changes and occlusions mainly since that the basic landmarks to initialize coarse alignment are not always available. Recently, local feature-based SIFT-like matching proves competent to handle all such variations without registration. In this paper, towards 3D face recognition for real-life biometric applications, we significantly extend the SIFT-like matching framework to mesh data and propose a novel approach using fine-grained matching of 3D keypoint descriptors. First, two principal curvature-based 3D keypoint detectors are provided, which can repeatedly identify complementary locations on a face scan where local curvatures are high. Then, a robust 3D local coordinate system is built at each keypoint, which allows extraction of pose-invariant features. Three keypoint descriptors, corresponding to three surface differential quantities, are designed, and their feature-level fusion is employed to comprehensively describe local shapes of detected keypoints. Finally, we propose a multi-task sparse representation based fine-grained matching algorithm, which accounts for the average reconstruction error of probe face descriptors sparsely represented by a large dictionary of gallery descriptors in identification. Our approach is evaluated on the Bosphorus database and achieves rank-one recognition rates of 96.56, 98.82, 91.14, and 99.21 % on the entire database, and the expression, pose, and occlusion subsets, respectively. To the best of our knowledge, these are the best results reported so far on this database. Additionally, good generalization ability is also exhibited by the experiments on the FRGC v2.0 database.

Keywords Registration-free 3D face recognition · Expression, pose and occlusion · 3D keypoint descriptors · Fine-grained matching

Communicated by C. Schnörr.

H. Li
School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, People's Republic of China

H. Li
Beijing Center for Mathematics and Information Interdisciplinary Sciences (BCMIIS), Beijing, People's Republic of China
e-mail: huibinli@mail.xjtu.edu.cn

D. Huang (✉) · Y. Wang
Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, Beijing 10091, People's Republic of China
e-mail: dhuang@buaa.edu.cn

Y. Wang
e-mail: yunhong@buaa.edu.cn

J.-M. Morvan
Département de Mathématiques, Université Claude Bernard Lyon 1, Lyon 69622, France

J.-M. Morvan
Geometric Modeling and Scientific Visualization Center, King Abdullah University of Science and Technology, Makkah, Saudi Arabia
e-mail: morvan@math.univ-lyon1.fr

L. Chen
Département de Mathématiques et Informatique, UMR CNRS 5205, Ecole Centrale Lyon, Lyon 69134, France
e-mail: liming.chen@ec-lyon.fr

1 Introduction

As it is natural, non-intrusive, and allows easy collection of face data, machine-based face recognition has received

Published online: 12 November 2014

significant attention from the biometrics community over the past several decades (Zhao et al. 2003). However, automatic face recognition in unconstrained environments without users' cooperation is currently a far more unsolved problem and a very challenging task (Li and Jain 2005). The main difficulties lie in the strong inter-subject similarities in facial appearance and the large intra-subject variations caused by severe illumination changes, large pose variations, partial occlusions and make up utilization (e.g. facial cosmetics) (Zhao et al. 2003; Li and Jain 2005).

Concerning these difficulties and along with the rapid development of 3D imaging systems, shape-based (3D) face recognition has been recently investigated as an alternative or complementary solution to traditional appearance-based (2D) face recognition. This research trend has been largely promoted by the release of benchmark databases like the Face Recognition Grand Challenge (FRGC v2.0) (Phillips et al. 2005), and a large number of 3D face recognition approaches have emerged in the past decade. See the early survey in Bowyer et al. (2006), and a thorough discussion in Spreeuwers (2011), Smeets et al. (2012), Huang et al. (2012), and Drira et al. (2013) for more recent contributions. However, the majority of these approaches are evaluated on face scans assumed to be captured with users' cooperation in constrained environments. In such a case, only frontal face scans with moderate expression variations are considered. Although high accuracies have been achieved, very sophisticated registration algorithms are usually indispensable (Kakadiaris et al. 2007; Faltemier et al. 2008; Al-Osaimi et al. 2009; Wang et al. 2010; Alyüz et al. 2010; Queirolo et al. 2010; Spreeuwers 2011; Mohammadzade and Hatzinakos 2013). More recently, 3D face recognition in real biometric applications using scans captured in less controlled or unconstrained conditions has received increasing interests (Passalis et al. 2011; Alyüz et al. 2013; Drira et al. 2013). In this scenario, 3D face recognition methods are expected to automatically and simultaneously deal with expression changes, occlusions, as well as pose variations. This directly leads to many uneasy issues such as automatic occlusion detection and restoration, pose normalization and fiducial point localization on partial face scans. All these difficulties suggest that developing 3D face recognition methods for real applications is a very challenging task.

1.1 Related Work

The literature does propose a few methods dealing with specific challenges of 3D face recognition in less controlled conditions. For example, in Passalis et al. (2011), facial symmetry is used to handle large pose variations for 3D face recognition in the real world; in Alyüz et al. (2013), a masked projection based on subspace analysis techniques is proposed for 3D face recognition under occlusions; in Drira et al. (2013), a curve-based shape analysis framework is presented for 3D face recognition under expression changes, occlusions and pose variations. However, all these methods require very sophisticated registration algorithms for automatic pose normalization (Passalis et al. 2011) or occlusion detection and restoration (Alyüz et al. 2013; Drira et al. 2013).

Fortunately, the well-known 2D SIFT matching framework and its 3D extensions, offer promising solutions to the above difficulties. The primary work of 3D SIFT-like matching framework performed on point cloud data is proposed in Mian et al. (2008), where 3D keypoints are first automatically determined on face scans by analyzing the differences between the principal axes of each local region. Then, a tensor representation based pose-invariant local shape descriptor is extracted at each keypoint. Finally, 2D SIFT matching under global graph constraint is used for keypoint matching. However, the tensor-based descriptor simply encodes the information of point coordinates and lacks of person dependent distinctiveness, thus limiting its performance in 3D face recognition. In Huang et al. (2012), facial range images are first represented by a group of extended local binary pattern (eLBP) maps, and the 2D SIFT matching framework is then performed on these images for keypoint detection, local feature extraction, and matching. Although this approach proves very discriminative and is registration-free for nearly frontal face scans, it cannot deal with large pose variations without manually labeled facial landmarks.

Considering the above limitations, it's necessary to extend the 2D SIFT matching framework to 3D mesh data. Since containing topological information, mesh data are more convenient for local operations (e.g. estimation of differential operators) than point cloud data, and thus have larger possibility to express more informative shape information, which directly leads to more discriminative local shape descriptors. Moreover, compared with 2.5D range images, pose-invariant local shape descriptors can be naturally constructed on full 3D free-form mesh data, and mesh-based 3D face recognition methods hence have the potential to simultaneously deal with expression changes, occlusions, as well as large pose variations. Recently, a few mesh data based SIFT-like matching methods have been proposed for 3D face recognition. However, their performance are still not sufficiently competitive (Li et al. 2011; Smeets et al. 2013; Berretti et al. 2013).

1.2 Contribution

Motivated by the main challenging issues and inspired by the 2D/3D SIFT methodology, we propose a mesh-based registration-free 3D face recognition approach that can uniformly and robustly deal with expression changes, occlusions, and pose variations. As shown in Fig. 1, our approach includes three basic modules: i.e. 3D keypoint detection,

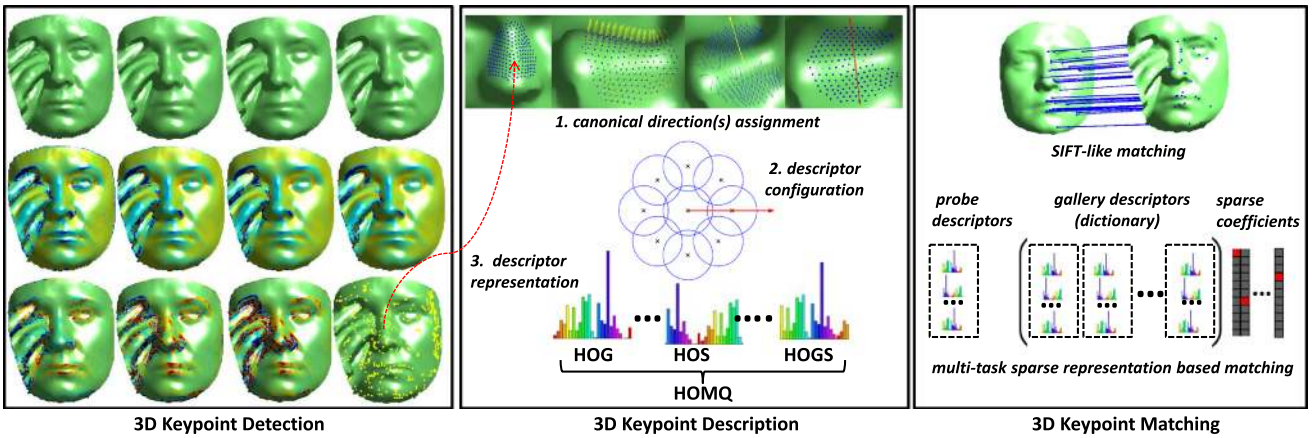

Fig. 1 Overview of the proposed method. 3D keypoint detection: from top to bottom, the original face scan and three smoothed face scans, their corresponding minimum principal curvature, Differences of Gaussian maps, and the detected 3D keypoints. The same procedures are also carried out for maximum principal curvature; 3D keypoint description: canonical direction assignment, quasi-daisy descriptor configuration, and descriptor representation by multi-order surface differential quantities; 3D keypoint matching: SIFT-like coarse-grained matching (*top*) and multi-task sparse representation-based fine-grained matching (*bottom*)

3D keypoint description, and 3D keypoint matching. Their details are presented as follows.

Our 3D keypoint detection algorithm starts by performing a sequence of Gaussian filters on a face scan, and continues by computing some scalar functions (e.g. curvatures) on the sequence of smoothed scans. The differences of Gaussian (DOG) operators are further computed in terms of these scalar functions and, finally, the local extrema of the DOG across scales are defined as keypoints. To locate more meaningful 3D keypoints, we propose using the two principal curvatures as the scalar functions.

For 3D keypoint description, to generate a descriptor with strong discriminative power, we propose building the weighted histograms of multiple order surface differential quantities, including the histograms of the surface gradient (1^{st} order), the surface shape index (2^{nd} order), the gradient of shape index (3^{rd} order), as well as their early fusion (multi-order). Our experimental results show that different orders of differential quantities have strong complementarity in descriptiveness, and their feature-level fusion can provide a comprehensive description of the 3D keypoint, thus ensuring very strong discriminative power.

For keypoint matching, instead of using the conventional SIFT matcher, which simply counts the pairs of corresponding keypoints, we propose a more precise matcher based on multi-task sparse representation. The proposed matcher first finds the sparsest representation of each probe descriptor from the complete dictionary set of all the descriptors associated with all the keypoints of the gallery scans. Then, the average reconstruction errors of sparse representation for the descriptor set, associated with all the keypoints of the probe scan, are computed as the similarity measurements between the probe scan and all the gallery scans. Finally, the gallery subject corresponding to the minimal error labels the identity of the probe.

Overall, our contributions involve all the above three modules and can be briefly summarized as follows.

(1) We introduce a principal curvature-based 3D keypoint detection algorithm, which can repeatedly identify complementary locations on the face scan where local curvatures are high (see Sect. 2).
(2) We present three novel pose-invariant 3D keypoint descriptors by computing the weighted histograms of different surface differential quantities, and their feature-level fusion for a comprehensive local shape description (see Sect. 3).
(3) A multi-task sparse representation-based fine-grained matching scheme is proposed, which makes use of the average sparse reconstruction error-based similarity measurement to significantly enlarge intra-subject similarity and reduce inter-subject similarity (see Sect. 4).
(4) We conduct comprehensive experiments on both the Bosphorus and FRGC v2.0 databases. Our approach achieves very competitive performance and shows good generalization ability (see Sects. 5 and 6).

It is worth distinguishing the main differences between the proposed approach and the very relevant work, meshSIFT (Maes et al. 2010; Smeets et al. 2013). On the keypoint detection module, both methods locate keypoints by finding local extrema of the DOG operator defined on curvature-based scale spaces. The meshSIFT algorithm uses the mean curvature to measure the saliency of keypoints, while we jointly use the two principal curvatures for saliency measurement. This strategy can largely increase the probability of locat-

ing more meaningful keypoints distributed at complementary facial positions and finding more informative geometry structures.

On the keypoint description module, the meshSIFT algorithm builds a histogram-based descriptor by combining the quantities of shape index and slant angle like Lo and Siebert (2009). In contrast, inspired by the daisy descriptor (Tola et al. 2010), we build four quasi-daisy histogram-based descriptors using three different surface differential quantities (mesh gradient, shape index, and the gradient of shape index) and their fusion. Experimental results based on the same setting show that our descriptor (the one with fusion) has much stronger discriminative power than meshSIFT.

On the keypoint matching module, the simple SIFT matcher is used by meshSIFT, while we propose a multi-task sparse representation based fine-grained matcher. The sparse representation based classifier was first introduced by Wright et al. (2009) for 2D face recognition. Recently, it has also been extended to expression-insensitive 3D face recognition (Li et al. 2009, 2014) and multi-pose 3D face recognition (Guo et al. 2013). In all these studies, the 2D face image or 3D face scan is generally represented by a single feature vector or multiple vectors with different scales. However, in our case, a 3D face scan is described by hundreds of unordered 3D keypoint descriptors: comparing two sets of these local descriptors using sparse representation is more complex. More recently, Liao et al. (2013) used multi-task sparse representation to compare two sets of local descriptors for 2D partial face recognition. Inspired by Liao et al. (2013); Wright et al. (2009), we build the fine-grained matcher for registration-free 3D face recognition. A similar matching scheme is also used for 3D-aided 2D face recognition, where the dictionary feature set is constructed by the different views of the gallery faces (Masi et al. 2013).

Preliminary results of our approach have been published in Li et al. (2011) and Veltkamp et al. (2011). However, significant extensions have been made since then, with notable improvement in the accuracy and robustness of the algorithm dealing with expression changes, occlusions, and pose variations. These extensions include testing the repeatability of the detected 3D keypoints, analyzing and comparing the discriminative power of different keypoint descriptors, using an SRC-based fine-grained matching scheme, performing detailed validation and comparisons on the entire Bosphorus database and its various subsets, conducting experiments on the entire FRGC v2.0 database, and evaluating time cost.

2 3D Keypoint Detection

Our 3D keypoint detection method is inspired by Lowe's SIFT (Lowe 2004) and some related work (Zaharescu et al. 2009, 2012; Maes et al. 2010; Smeets et al. 2013), but with targeted improvements. The details are introduced as follows.

2.1 Principal Curvature-Based 3D Keypoint Detectors

(i) *Scale-Space Construction* Keypoint detection starts by performing a sequence of Gaussian filters G_{σ_s} with different scales σ_s over a face mesh S. For a generic point $\mathbf{p} \in S$, the Gaussian filter modifies the geometry structure of S by updating the coordinates of $\mathbf{p}$ as follows:

$$\mathbf{p}_s = \frac{\sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p},1)} w_s(\mathbf{p},\mathbf{q})\cdot\mathbf{q}}{\sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p},1)} w_s(\mathbf{p},\mathbf{q})}, \tag{1}$$

where $\mathbf{p}_s$ represents the updated point at scale σ_s, $\mathcal{N}(\mathbf{p}, 1)$ represents the point set of the one-ring neighbor of $\mathbf{p}$, and

$$w_s(\mathbf{p},\mathbf{q}) = G_{\sigma_s}(d_e(\mathbf{p},\mathbf{q})) = \exp(-\|\mathbf{p}-\mathbf{q}\|^2/2{\sigma_s}^2). \tag{2}$$

(ii) *Scale-Space Extrema* Once the scale-space is constructed, a scalar function f, used for measuring the saliency of a keypoint, is computed for each point at each scale. In theory, this function could be any scalar function $f(\mathbf{p}) : S \to \mathcal{R}$ defined on a face mesh. However, 3D face scans are discrete approximations of facial surfaces and surface curvatures, the fundamental differential geometry quantities for the description of local shapes, naturally come into sight. In order to detect more meaningful 3D keypoints located around areas with high local curvatures, we propose to use both of the two principal curvatures as the scalar functions. According to Goldfeather and Interrante (2004), the principal curvatures are computed by fitting a cubic-order surface patch:

$$f(x,y) = \frac{A}{2}x^2 + Bxy + \frac{C}{2}y^2 + Dx^3 + Ex^2y + Fxy^2 + Gy^3 \tag{3}$$

and its normal vectors $(f_x(x, y), f_y(x, y), -1)$ using both the 3D coordinates and the normal vectors of the associated local neighbor points (two-ring). The maximum principal curvatures $C_M(\mathbf{p})$ and the minimum principal curvatures $C_m(\mathbf{p})$ at a given point $\mathbf{p}$ are computed as the eigenvalues of the *Weingarten* matrix. With the choice of two principal curvatures, the differences of Gaussian operators (approximations of the Laplacian operators) are then defined as follows:

$$\rho(C_M, \mathbf{p}_s) = C_M(\mathbf{p}_s) - C_M(\mathbf{p}_{s-1}), \tag{4}$$

$$\rho(C_m, \mathbf{p}_s) = C_m(\mathbf{p}_s) - C_m(\mathbf{p}_{s-1}). \tag{5}$$

3D keypoints are detected by finding the extrema of $\rho(C_M, \mathbf{p}_s)$ and $\rho(C_m, \mathbf{p}_s)$ across scales using a one-ring neighborhood of each point, respectively.

As mentioned above, the mean curvature function $C_H(\mathbf{p})$ is used as the scalar function in the meshSIFT algorithm

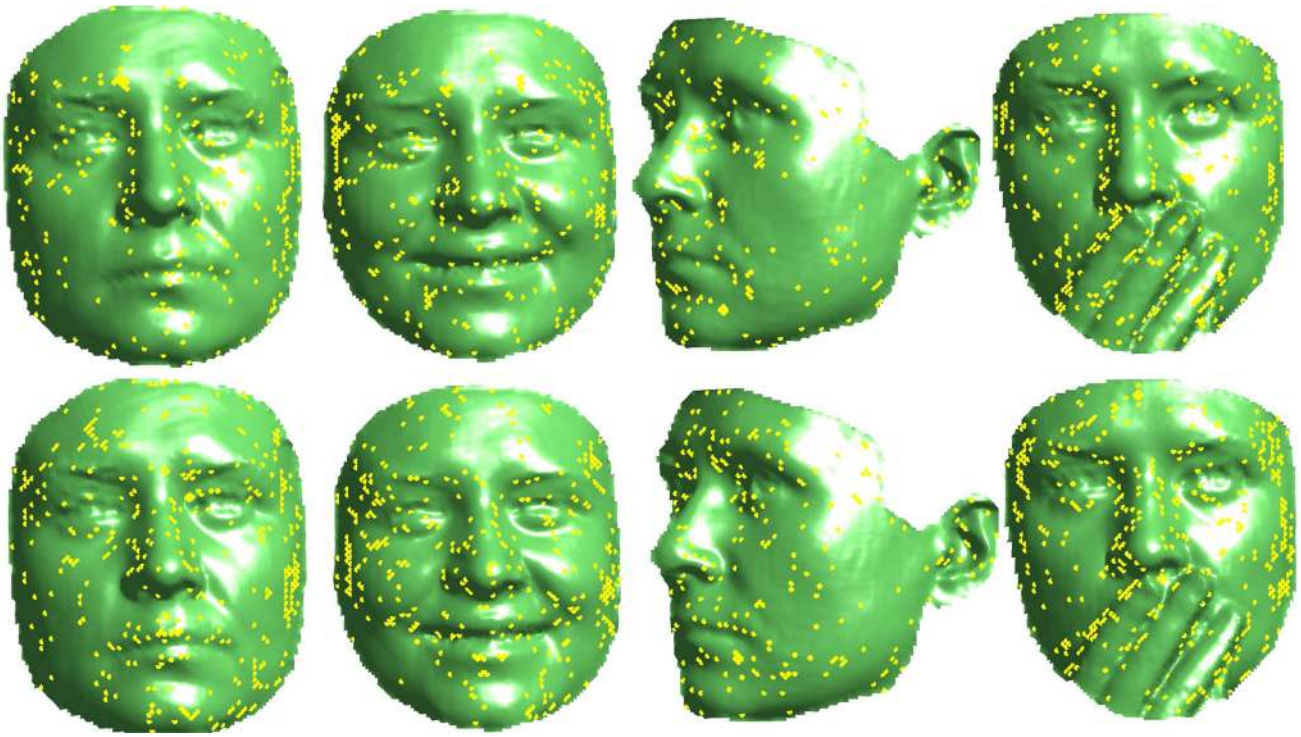

Fig. 2 The detected 3D keypoints by C_M (*top*) and C_m (*bottom*) for a neutral face, a happy face, a 45° rotated face, and a face with mouth occlusion

(Maes et al. 2010; Smeets et al. 2013). In this paper, we will show that the joint use of two principal curvatures can locate more meaningful 3D keypoints, and thus achieve much better identification performance than the mean curvature.

2.2 Keypoint Distribution

As shown in Fig. 2, for both principal curvatures (C_M and C_m) in the faces with different expressions, poses, and occlusions, the keypoints mainly locate in the eye, nose, and mouth (or occluded mouth) regions, and sparsely distribute over other regions like the forehead, cheekbone, and chin. In general, the principal curvature-based detectors tend to extract keypoints located around areas characterized by high local curvatures. We can see that the keypoints detected by C_M and C_m distribute at many complementary positions, such as the lip, nasal, and eye regions. Our main idea is that joint use of two principal curvatures increases the probability of locating more meaningful facial points with different shape structures, such as the *elliptic* points ($C_M(\mathbf{p}) \cdot C_m(\mathbf{p}) > 0$), *hyperbolic* points ($C_M(\mathbf{p}) \cdot C_m(\mathbf{p}) < 0$) as well as *parabolic* points ($C_M(\mathbf{p}) \cdot C_m(\mathbf{p}) = 0$). Moreover, the keypoints sparsely distribute over the entire facial local regions (rigid or non-rigid, frontal or rotated, occluded or non-occluded), which provides the possibility of dealing with the expression, pose and occlusion problems at a unified framework.

2.3 Keypoint Repeatability

Besides the keypoint distribution, repeatability is another main feature of a 3D keypoint detector (Tombari et al. 2013). Given two arbitrary face scans of the same subject, repeatability of detected 3D keypoints implies that the two sets of keypoints extracted from the two face scans at the same pose are roughly located at the same position. However, there is no ground truth to check such repeatability over 3D face scans of the same subject. In this paper, we adopt the same method as in Mian et al. (2008) and evaluate on the Bosphrous database. As shown in Fig. 3, the repeatability of C_M(-Max) and C_m(-Min) reaches 68 % (61 %, resp.) and 75 % (68 %, resp.) respectively for faces with neutral expression (non-neutral, resp.) at an error of 6 mm. This is expected as the degradation between neutral samples and non-neutral ones is mainly due to the fact that facial expressions elastically deform surfaces and thereby induce different keypoints. However, repeatability in such a case still remains as high as 68 % for C_m (61 % for C_M, resp.), indicating that much information within local regions around keypoints can be used for matching.

3 3D Keypoint Description

Globally, distribution and repeatability of 3D keypoints are crucial as introduced in the previous sections; while, locally,

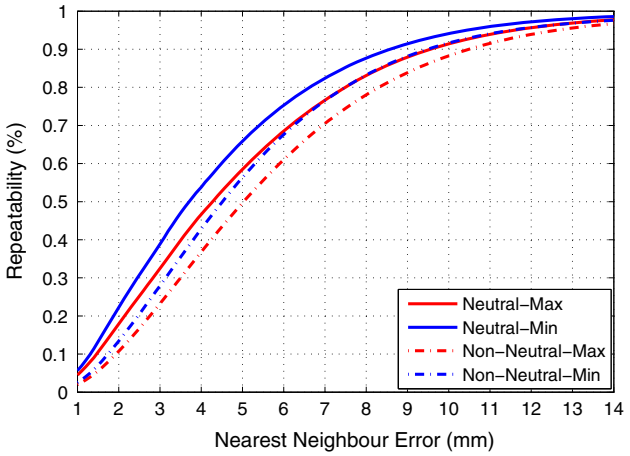

Fig. 3 Repeatability of 3D keypoints on the Bosphorus database

the discriminative power of each 3D keypoint representation (i.e. keypoint descriptor) is also crucial for the following face matching and recognition.

Based on this consideration, we propose three keypoint descriptors, namely the Histogram of Gradient (HOG), the Histogram of Shape index (HOS), and the Histogram of Gradient of Shape index (HOGS). Intuitively, HOG describes the point-level bending pattern of the local shape around a keypoint; HOS indicates the distribution of different shape categories; and HOGS depicts the changing pattern of different shape categories. These three descriptors comprise different orders of surface differential quantities, and thus have strong complementarity in descriptiveness. We further build a more comprehensive keypoint descriptor, namely the Histogram of Multiple surface differential Quantities (HOMQ) by combining HOG, HOS, and HOGS at feature level.

Our proposed descriptor, i.e. HOMQ, is similar in spirit to the 2D SIFT (Lowe 2004), 2.5D SIFT (Lo and Siebert 2009), meshHOG (Zaharescu et al. 2009, 2012), and meshSIFT (Maes et al. 2010; Smeets et al. 2013), but encodes more shape characteristics, since it includes the histograms of three different surface differential quantities. The details are introduced as follows.

(i) *Canonical direction assignment* The local descriptors for keypoint $\mathbf{p} \in S$ are computed within a geodesic disc of radius R:

$$\mathcal{N}(\mathbf{p}) = \{\mathbf{q} | d_g(\mathbf{p}, \mathbf{q}) < R\}, \tag{6}$$

where $d_g(\mathbf{p}, \mathbf{q})$ denotes the geodesic distance between $\mathbf{p}$ and $\mathbf{q}$ (computed by the *Toolbox Fast Marching*). To achieve rotation invariance, a canonical direction (see Fig. 4) is necessary for each keypoint, which can be robustly assigned based on a translation and rotation invariant local coordinate system as in Skellya and Sclaroffb (2007). First, all points $\mathbf{q} \in \mathcal{N}(\mathbf{p})$ and their normal vectors $\mathbf{n}(\mathbf{q})$ are transformed to the following temporary local coordinate system:

$$C = \{\mathbf{t}(\mathbf{p}'), \mathbf{t}(\mathbf{p}') \times \mathbf{n}(\mathbf{p}'), \mathbf{n}(\mathbf{p}')\}, \tag{7}$$

where $\mathbf{p}'$ (transformed point of $\mathbf{p}$) is the origin, its unit normal $\mathbf{n}(\mathbf{p}')$ is the z-axis, and $\mathbf{t}(\mathbf{p}')$ is the x-axis, which is a randomly selected initial unit vector in the tangent plane $\mathcal{T}_{\mathbf{p}'}$ of surface S at $\mathbf{p}'$. Then, the transformed points $\mathbf{q}'$ and their normal vectors $\mathbf{n}(\mathbf{q}')$ are projected to the tangent plane $\mathcal{T}_{\mathbf{p}'}$. Their gradients $\theta(\mathbf{q}')$ and corresponding magnitudes $mag(\mathbf{q}')$ are computed as:

$$\theta(\mathbf{q}') = \arctan[\mathbf{n}_y(\mathbf{q}')/\mathbf{n}_x(\mathbf{q}')], \tag{8}$$

$$mag(\mathbf{q}') = \sqrt{\mathbf{n}_x^2(\mathbf{q}') + \mathbf{n}_y^2(\mathbf{q}')}, \tag{9}$$

where $\mathbf{n}_x(\mathbf{q}') = \mathbf{t}(\mathbf{p}') \cdot \mathbf{n}(\mathbf{q}')$, $\mathbf{n}_y(\mathbf{q}') = \mathbf{t}(\mathbf{p}') \times \mathbf{n}(\mathbf{p}') \cdot \mathbf{n}(\mathbf{q}')$. Here, $\mathbf{n}(\cdot)$ is simply computed by averaging the triangles' normals within the one-ring neighborhood of the associated point.

Finally, a Gaussian weighted gradient histogram of 360 bins (1 bin per degree) is constructed on the tangent plane $\mathcal{T}_{\mathbf{p}'}$ of the temporary local coordinate system. The Gaussian (see (2)) weights are represented as:

$$w(\mathbf{p}', \mathbf{q}') = mag(\mathbf{q}') \cdot G_\sigma(d_g(\mathbf{p}', \mathbf{q}')), \tag{10}$$

where the standard deviation σ is set to half of the radius R. Similar to the SIFT descriptor (Lowe 2004), the canonical

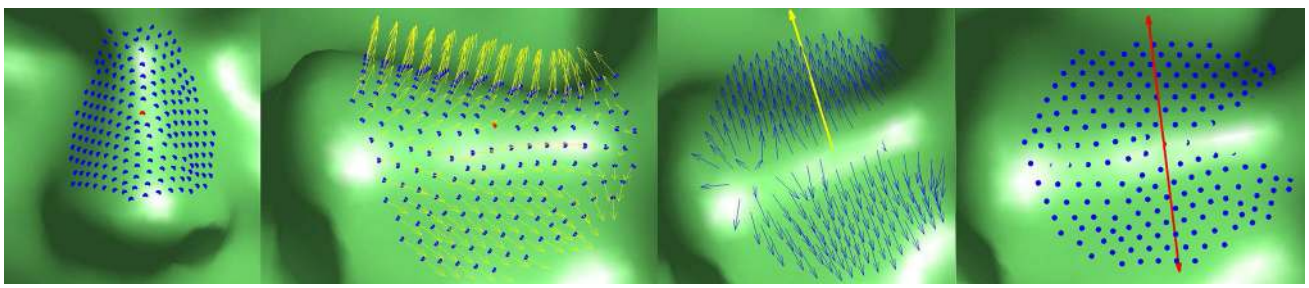

Fig. 4 Canonical direction assignment. From *left* to *right* a detected keypoint (*in red*) and its geodesic disk patch points (*in blue*); normal vectors (*in yellow*) of the associated points; projected normal vectors and the initial direction (*in yellow*); projected points and two assigned canonical directions (*in red*) (Color figure online)

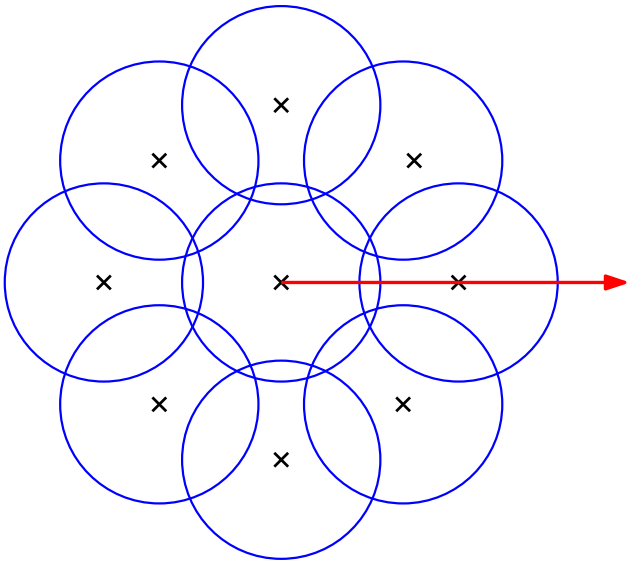

Fig. 5 The quasi-daisy spatial configuration.The *cross* sign represents a 3D keypoints and its 8 neighboring points. Each *circle* represents a region where we compute the descriptor. By overlapping the regions we achieve smooth shifting between regions. The 'red vector' denotes the canonical direction for rotation invariance

direction $\mathbf{d}(\mathbf{p}')$ of $\mathbf{p}'$ is assigned by the peak of the weighted gradient histogram. Noting that more than one canonical direction may be assigned to a keypoint, for simplicity, we assume that only one canonical direction can be assigned to each keypoint in the subsequent. Once $\mathbf{d}(\mathbf{p}')$ is assigned, all the neighbor points are transformed to a new local coordinate system:

$$C' = \{\mathbf{d}(\mathbf{p}'), \mathbf{d}(\mathbf{p}') \times \mathbf{n}(\mathbf{p}'), \mathbf{n}(\mathbf{p}')\} \tag{11}$$

for the following processes.

(ii) *Descriptor Configuration* Descriptor configuration is performed on the tangent plane $\mathcal{T}_{\mathbf{p}'}$ of the new local coordinate system C'. Inspired by the 2D daisy descriptor (Tola et al. 2010), 9 overlapping circular regions with a radius of r_2 are assigned centering at the keypoint $\mathbf{p}'$ and its 8 neighboring points, respectively (see Fig. 5). This kind of quasi-daisy radial flower pattern of overlapping circles simulates the functioning of human complex cells in the visual cortex (Hubel and Wiesel 1962). Therefore, it tends to be robust to small transformations, e.g. spatial shifting, non-rigid deformations. The 8 neighboring points are localized by performing uniform sampling over a circle centered at $\mathbf{p}'$ with a radius of r_1, starting from the canonical direction $\mathbf{d}(\mathbf{p}')$.

(iii) *Descriptor Representation* In each circular region c_i $(i = 1, 2, \cdots, 9)$, we construct the local weighted histograms of different surface differential quantities, including the values of surface gradient (see (9)), shape index, as well as the gradient of shape index. The shape index $SI(\mathbf{p})$ at point $\mathbf{p}$ can be computed based on its two principal curvatures as follows,

$$SI(\mathbf{p}) = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{C_M(\mathbf{p}) + C_m(\mathbf{p})}{C_M(\mathbf{p}) - C_m(\mathbf{p})}. \tag{12}$$

According to Meyer et al. (2001)), the gradient of the shape index function $\nabla SI(\mathbf{p})$ can be computed by solving the following optimization problem:

$$\arg \min_{\nabla SI(\mathbf{p})} \sum_{\mathbf{q} \in \mathcal{N}(\mathbf{p},1)} |\nabla SI(\mathbf{p})^T proj(\overrightarrow{\mathbf{pq}}) - \frac{SI(\mathbf{p}) - SI(\mathbf{q})}{\|\mathbf{p} - \mathbf{q}\|}|, \tag{13}$$

where $proj(\overrightarrow{\mathbf{pq}})$ represents a vector computed by projecting the unit vector $\overrightarrow{\mathbf{pq}}$ to the tangent plane $\mathcal{T}_{\mathbf{p}}$. Their corresponding histograms are referred to as the histogram of gradient (hog_i), histogram of shape index (hos_i) and histogram of gradient of shape index $(hogs_i)$, respectively. For hog_i and $hogs_i$, the histograms of gradient angles with 8 bins representing 8 main orientations ranging from 0° to 360° are computed and weighted by their corresponding gradient magnitudes. For hos_i, the shape index values ranging from 0 to 1 are also equally quantized to 8 bins, and weighted by a Gaussian kernel G_σ, where the standard deviation is set as the Euclidian distance between the current point and the center point of the circle region. The final histograms at keypoint $\mathbf{p}$ are constructed by concatenating hog_i, hos_i and $hogs_i$ in a clockwise direction, represented as:

$$\text{HOG} = (hog_1, hog_2, \ldots, hog_9), \tag{14}$$

$$\text{HOS} = (hos_1, hos_2, \ldots, hos_9), \tag{15}$$

$$\text{HOGS} = (hogs_1, hogs_2, \ldots, hogs_9). \tag{16}$$

The above sub-histograms (e.g. hog_i) and histograms (e.g. HOG) are all normalized to unit vectors to eliminate the influence of non-uniform mesh sampling. This generates three 3D keypoint descriptors with the same dimension of 72. As mentioned above, these three local descriptors contain strong complementarity information in descriptiveness. Finally, we build a more comprehensive and discriminative keypoint descriptor HOMQ, which is obtained by feature-level fusion of the above three descriptors, and thus has a dimension of 216.

4 3D Keypoint Matching

According to the framework of keypoint detection, description and matching, the most direct similarity measurement between a pair of probe-gallery face shapes is the total number of their corresponding keypoints. Here, the one-to-one keypoint correspondence can be established using the classical SIFT matcher proposed by Lowe (2004). The SIFT matcher uses the angle as the similarity measurement of descriptors. The smaller the angle, the more similar the descriptors and vice versa. A match is accepted if the ratio

between the *arcos* distance of the best match (i.e. the minimal angle) and the second best match is lower than a threshold μ. Due to its simplicity, the SIFT matcher has been widely used in 3D face recognition such as Mian et al. (2007, 2008), Maes et al. (2010), Li et al. (2011), Huang et al. (2012), Smeets et al. (2013) and so on. To improve its robustness, holistic spatial constraints or the RANdom SAmple Consensus (RANSAC) algorithm are commonly used, see Mian et al. (2008), Huang et al. (2012), Smeets et al. (2013) and Berretti et al. (2013) for the specific details.

However, the number of corresponding keypoints in nature is a coarse-grained similarity measurement, and the SIFT matcher has thus proved sensitive to missing data (e.g. caused by large pose variations of face scans). Based on this consideration and inspired by the Sparse Representation based Classifier (SRC) (Wright et al. 2009; Liao et al. 2013), we propose a new SRC-based matching scheme. A similar SRC-based matcher has also been developed in Masi et al. (2013) for 3D-aided 2D face recognition in the wild. In comparison with the SIFT matcher, which coarsely counts the number of matched keypoints, the SRC-based matcher precisely computes the normalized (average) accumulative sparse reconstruction error for all the keypoints of a probe face. For this reason, we call the SRC-based matcher as the Fine-Grained Matcher (FGM) in the subsequent. In contrast, the SIFT matcher is subsequently denoted as the Coarse-Grained Matcher (CGM). The FGM includes two main procedures: the construction of a gallery dictionary and the computation of sparse reconstruction errors, which are introduced as follows.

4.1 Gallery Dictionary Construction

Given a gallery set of N subjects, each subject has a single 3D face scan. Assume that n_i 3D keypoints are detected for i-th subject. Let the corresponding n_i 3D keypoint descriptors be the following sub-dictionary:

$$\mathbf{D}_i = [\mathrm{d}_{i,1}, \mathrm{d}_{i,2}, \cdots, \mathrm{d}_{i,n_i}] \in \mathcal{R}^{m \times n_i}, \tag{17}$$

where m represents the descriptor dimension. In our case, $m = 72$ for HOG, HOS, and HOGS; $m = 216$ for HOMQ. Then, the dictionary for all the N subjects of the gallery can be built by simply concatenating all the sub-dictionaries as:

$$\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_N] \in \mathcal{R}^{m \times K}, \tag{18}$$

where $K = n_1 + n_2 + \cdots + n_N$ represents the total number of keypoint descriptors in the gallery. In practice, since hundreds of keypoints can be detected for each 3D face scan, K is very large, making $\mathbf{D}$ an over-complete dictionary containing informative local shape atoms of all the N subjects.

4.2 Multi-task Sparse Representation

Given a probe face scan with n 3D keypoint descriptors

$$\mathbf{Y} = (\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_n). \tag{19}$$

Then multi-task sparse representation of $\mathbf{Y}$ by $\mathbf{D}$ can be formulated as:

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}} \sum_{i=1}^{n} \|\mathrm{x}_i\|_0, \quad s.t. \ \mathbf{Y} = \mathbf{D}\mathbf{X}, \tag{20}$$

where $\mathbf{X} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n) \in \mathcal{R}^{K \times n}$ is the sparse coefficient matrix, and $\|\cdot\|_0$ denotes the l_0 norm of a vector, defined as the number of non-zero elements of the vector. As the probe descriptors are independent from each other, we can equivalently solve the following nl_0-minimization problems:

$$\hat{\mathrm{x}}_i = \arg\min_{\mathrm{x}_i} \|\mathrm{y}_i - \mathbf{D}\mathrm{x}_i\|_2 \ \ s.t. \ \|\mathrm{x}_i\|_0 \le L, \quad i = 1, 2, \cdots, n, \tag{21}$$

where L is the sparsity parameter, which controls the sparsity of the solution. To solve (21), we make use of the Orthogonal Matching Pursuit (OMP) algorithm proposed by Pati et al. (1993). Inspired by Wright et al. (2009), we introduce the following multi-task SRC to determine the identity of the probe face scan,

$$identity(\mathbf{Y}) = \arg\min_{j} \frac{1}{n} \sum_{i=1}^{n} \|\mathrm{y}_i - \mathbf{D}\delta_j(\hat{\mathrm{x}}_i)\|_2^2, \tag{22}$$

where $\delta_j(\cdot)$ is a characteristic function which selects only the coefficients associated with the j-th subject. As mentioned above, (22) computes the normalized (average) accumulative sparse reconstruction error for all the keypoints of a given probe face scan.

Intuitively, if a probe face scan and a gallery face scan belong to the same identity, they should share a large set of similar keypoints whose local geometric characteristics are also similar to each other. Thus, for any descriptor of the probe face, there is a high probability of selecting the descriptors of the gallery face. That means that most of the errors are close to 0, and only a few are equal to 1. So, the accumulative sparse reconstruction error will still be very small. In this case, the probe face can be probably recognized.

5 Performance Evaluation

5.1 Database

We mainly evaluate the effectiveness of the proposed 3D face recognition approach on the Bosphorus database (Savran et

al. 2008). It consists of 4,666 3D face scans of 105 subjects made up of 61 men and 44 women. For each subject, there are around 34 expressions, 13 poses, and 4 occlusions. According to different variations, we divide the entire database into three subsets: (i) *Expressions* This subset contains 6 basic expressions (Anger, Disgust, Fear, Happiness, Sadness, and Surprise) along with Neutral; 28 facial Action Units (AUs): 20 Lower AUs (LAU), 5 Upper AUs (UAU), 3 Combined AUs (CAU). (ii) *Poses* This subset includes 7 Yaw Rotations (YR): $+10°$, $+20°$, $+30°$, $\pm 45°$, $\pm 90°$; 4 Pitch Rotations (PR): strong upwards/downwards, slight upwards /downwards; 2 Cross Rotations (CR): $+45°$ yaw and approximately ± 20 pitch. (iii) *Occlusions* This subset comprises occlusions of eyes by hand (E-Hand), mouth by hand (M-Hand), eyes by glasses (E-Glasse), and facial regions by hair (F-Hair), e.g. long hair for females and facial hair like beards and moustaches for males. Moreover, a subset of 18 unlabeled frontal expression scans without any occlusions is included. The 3D face scans are acquired using structure-light based 3D scanning system. The sensor resolution in x, y, and z (depth) dimensions are 0.3 mm, 0.3 mm, and 0.4 mm respectively. After preprocessing by the providers, facial regions without clutter are cropped from the raw data, and each face scan approximately contains 35,000 points. In our experiments, all the face scans are first down-sampled and then triangulated by simply connecting the neighboring vertices. For performance evaluation, the first neutral scan of each subject is used to construct the gallery, and the remaining scans or their subsets are used as probes.

5.2 Parameters

In our experiments, we empirically choose the parameters of each module as follows. For 3D keypoint detection, 3 scales are used, and the scale parameters σ_s are set as 1.83, 2.50, and 4.80 mm, respectively. For 3D keypoint description, we set the radius R of the geodesic disc to 22.5 mm. The parameters for descriptor configuration r_1 and r_2 are set to 15 and 10 mm, respectively. For 3D keypoint matching, the SIFT matching ratio μ for CGM is set at 0.70 for the HOG descriptor, and at 0.75 for the other descriptors. Moreover, the sparsity parameter L for FGM is set to 1. That is to say, we seek the sparsest representation of a probe descriptor among all the corresponding gallery descriptors. However, the matching algorithm is not sensitive to this parameter, and small changes in the value (e.g. $L = 5$) reach similar performance.

5.3 Performance

Table 1 reports rank-one recognition rates of the proposed approach on the entire Bosphorus database and its various subsets, containing the performance comparison of the proposed descriptors (HOG, HOS, HOGS, and HOMQ) and matchers (CGM and FGM). All the results are achieved by performing score-level fusion of C_M and C_m based 3D keypoint detectors. For CGM, the sum of their matched numbers is used for the similarity measurement, whereas for FGM the sum of their normalized reconstruction errors is adopted. We can interpret these results from the following three aspects:

(i) *Discriminability of descriptors* in terms of the discriminative power of descriptors, HOMQ generally performs better than the others over two matchers and across all face variations (subsets). This indicates that HOMQ, which fuses three different kinds of surface differential quantities, can provide a comprehensive description of local shape, and thus has the strongest discriminative power. Moreover, HOS performs much better than HOG and HOGS over CGM matcher and across all face variations, especially for the pose subset (e.g. YR45°). However, this superiority becomes unconspicuous over FGM matcher, and the difference is compensated by the advantage of FGM. We can therefore conclude that HOS has stronger discriminative power than HOG and HOGS, while HOG and HOGS are comparable to each other.

(ii) *Effectiveness of Matchers* concerning the effectiveness of matchers, FGM clearly outperforms CGM over all descriptors and across all subsets. This advantage is very significant for the HOG and HOGS descriptors, and, in both cases, more than 10 % improvements are achieved over the entire Bosphorus database. Meanwhile, this superiority becomes more significant over subsets with large pose variations. For example, there are more than 40 % improvements on the YR45° subset for HOG and the CR subset for HOGS. Moreover, the performance improvements for HOS and HOMQ descriptors are particularly significant on some very difficult subsets. For example, there are more than 10 % improvements on the Anger, Disgust, Happy, YR45° and CR subsets for HOS; and on the Disgust and YR90° subsets for HOMQ. All these results prove that the SRC-based fine-grained matcher (FGM) is more efficient than the SIFT-like coarse-grained matcher (CGM). It indicates that, the finer the similarity measurement, the stronger the matcher. It is worth noting that FGM is not sensitive to descriptors. For example, HOG, HOS, and HOGS achieve very similar results on many subsets although they have different discriminative powers as pointed out above. As shown in Yang et al. (2007), the reason is if sparsity in the recognition problem is properly harnessed, the choice of features is no longer critical. What is critical, however, is whether the number of features is sufficient and whether the sparse representation is correctly found.

(iii) *Robustness to Variations* regarding the robustness of our approach to different variations, we consider the combination of the HOMQ descriptor and the FGM matcher. On the expression subset, our approach achieves very competitive rank-one recognition rates, and even 100 % on the Neutral, Sad, UAU, and CAU subsets. On the pose subset, most of the accuracies are more than 97 %, and even 100 %

Table 1 Performance in terms of rank-one recognition rates on the subsets of expressions, poses, occlusions, unlabeled, and the entire Bosphorus database

	HOG (%)		HOS (%)		HOGS (%)		HOMQ (%)	
	CGM	FGM	CGM	FGM	CGM	FGM	CGM	FGM
Neutral (105 scans) vs. expressions (2,797 scans)								
Neutral (194)	99.48	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Anger (71)	69.01	91.55	87.32	98.59	76.06	94.37	88.73	97.18
Disgust (69)	50.72	84.06	56.52	78.26	60.87	88.41	76.81	86.96
Fear (70)	71.43	92.86	88.57	92.86	84.29	94.29	92.86	98.57
Happiness (106)	79.25	91.51	85.85	96.23	75.47	98.11	95.28	98.11
Sadness (66)	80.30	95.45	90.91	96.97	86.36	96.97	95.45	100.0
Surprise (71)	81.69	100.0	97.18	100.0	84.51	97.18	98.59	98.59
LAU (1,549)	88.96	96.97	95.09	98.13	90.70	98.52	97.22	98.84
UAU (432)	94.21	99.07	98.15	100.0	95.14	99.54	99.07	100.0
CAU (169)	92.90	100.0	97.04	99.41	95.86	100.0	98.82	100.0
All (2,797)	88.09	96.96	94.32	97.96	90.24	98.32	96.89	98.82
Neutral (105 scans) vs. poses (1,365 scans)								
YR10 (105)	98.10	100.0	100.0	100.0	96.19	100.0	100.0	100.0
YR20 (105)	90.48	99.05	96.19	99.05	87.62	98.10	99.05	100.0
YR30 (105)	84.76	98.10	92.38	100.0	75.24	96.19	98.10	99.05
YR45 (210)	49.52	89.05	80.48	91.90	49.05	84.76	90.95	97.62
YR90 (210)	07.62	16.67	17.62	25.24	02.86	10.00	33.33	47.14
YR (735)	55.37	72.65	69.25	76.19	51.84	71.43	77.96	84.08
PR (419)	94.27	99.28	96.90	98.57	84.73	99.28	98.81	99.52
CR (211)	60.66	90.05	79.62	94.79	48.34	89.10	94.31	99.05
All (1,365)	68.13	83.52	79.34	85.93	61.39	81.47	86.89	91.14
Neutral (105 scans) vs. occlusions (381 scans)								
E-Hand (105)	96.19	100.0	100.0	100.0	96.19	100.0	100.0	100.0
M-Hand (105)	90.48	100.0	97.14	98.10	93.33	99.05	99.05	100.0
E-Glasses (104)	95.19	100.0	97.12	97.12	97.12	97.12	100.0	100.0
F-Hair (67)	86.57	94.03	94.03	91.04	86.57	82.09	97.01	95.52
All (381)	92.65	98.95	97.38	97.90	93.96	96.59	99.21	99.21
Neutral (105 scans) vs. unlabeled scans (18 scans)								
All (18)	88.89	100.0	100.0	100.0	94.44	100.0	100.0	100.0
Neutral (105 scans) vs. all meshes except yaw 90 (4,351 scans)								
All (4,351)	86.12	96.81	93.61	97.70	85.75	97.15	97.04	98.94
Neutral (105 scans) vs. all scans (4,561 scans)								
All (4,561)	82.50	93.12	90.11	94.37	81.93	93.14	94.10	96.56

Four proposed keypoint descriptors (HOG, HOS, HOGS, and HOMQ) combined with two keypoint matchers (CGM and FGM) are compared with each other. The names and the number of probe scans for each subset are listed in the left-hand column

on the YR10° and YR20° subsets. The most challenging subset is YR90°, where only half of the face scan (left or right profile) is available. In this case, our framework can still correctly recognize about half of the face scans (47.14 %). On the occlusion subset, our approach can correctly classify all the probes, except for three which are heavily occluded by hair. All these results consistently demonstrate that the proposed approach is very robust for 3D face recognition under expression changes, occlusions, and pose variations.

5.4 Comparison

Table 2 shows the performance comparison between our approach (HOMQ+FGM) and the state-of-the-art ones on the Bosphorus database and its various subsets. For a fair comparison, the same experimental protocol of Table 1 is used. It is worth noting that the probe subsets differ slightly according to the method. For the expression subset, 2,797 expression scans plus 18 unlabeled scans are used in Alyüz

Table 2 Comparison of rank-one recognition rates on the subsets of expressions, poses, occlusions, and the entire Bosphorus database

	Expressions (%)	Poses (%)	YR (%)	YR90° (%)	PR (%)	CR (%)	Occlusions (%)	E-Hand (%)	M-Hand (%)	E-Glasses (%)	F-Hair (%)	All scans (%)
Alyüz et al. (2008)	–	–	–	–	–	–	93.6	93.6	93.6	97.8	89.6	–
Alyüz et al. (2010)	98.2	–	–	–	–	–	–	–	–	–	–	–
Colombo et al. (2011)	–	–	–	–	–	–	87.6	91.1	74.7	94.2	90.4	–
Ocegueda et al. (2011)	98.2	–	–	–	–	–	–	–	–	–	–	–
Drira et al. (2013)	–	–	–	–	–	–	87.0	97.1	78.0	94.2	81.0	–
Smeets et al. (2013)	97.7	84.2	–	24.3	–	–	–	–	–	–	–	93.7
Berretti et al. (2013)	95.7	88.6	81.6	45.7	98.3	93.4	93.2	–	–	–	–	93.4
This paper	98.8	91.1	84.1	47.1	99.5	99.1	99.2	100	100	100	95.5	96.6

et al. (2010); and 3,186 frontal scans are used in Smeets et al. (2013). For the occlusion subset, 360 scans rather than all the 381 occluded probe scans are used in Colombo et al. (2011).

From Table 2, we can see that our proposed method significantly outperforms all the other methods, especially on the subsets of occlusions and pose variations, and achieves the best rank-one recognition rates on the entire database and these defined subsets. It should be noted that the high performance in Alyüz et al. (2010) and Ocegueda et al. (2011) achieved on the expression subset relies heavily on sophisticated face registration algorithms. Similarly, the registration-based methods Alyüz et al. (2008), Colombo et al. (2011), and Drira et al. (2013) deal with the occlusion problem with the help of additional face data and subspace learning techniques such as PCA (Principal Component Analysis). In contrast, the registration-free methods in Smeets et al. (2013), Berretti et al. (2013) and this paper, which are based on the SIFT-like framework, can simultaneously handle facial expression changes, occlusions, and pose variations.

To sum up, from Table 2, we can conclude that, as a result of capturing more meaningful keypoints, extracting more discriminative descriptors, and building more efficient matcher, the method we propose has proved more effective and robust than the high related counterpart, i.e. Smeets et al. (2013), Berretti et al. (2013) although they derive from the same framework. Moreover, our result on the YR90° subset indicates that this scenario is still a very challenging issue in 3D face recognition.

6 Discussion

6.1 3D Keypoint Detectors: Choice of Scalar Function

As mentioned in Sect. 2, in theory, the scalar function, defined as $f(\mathbf{p}) : S \rightarrow \mathcal{R}$ on the mesh, of the scale-space used for keypoint detection, has various choices, and is expected to be able to capture some structure information on the mesh. Thus, surface curvatures, as the most important differential geometry quantities for shape characterization, become the main a major alternative in the literature. An interesting question to ask is which curvature is better? The mean curvature is used in the meshSIFT algorithm (Maes et al. 2010; Smeets et al. 2013) and the meshHOG based 3D face recognition algorithm (Berretti et al. 2013). However, in this study, we suggest using both the maximum and minimum principal curvatures instead of the mean curvature. This is mainly due to the fact that, for a given surface, the combination of two principal curvatures can provide more complete local shape characterization, and thus find more meaningful 3D keypoints.

The results shown in Table 3 support the above conclusion. In Table 3, different curvatures are used as the scalar func-

Table 3 Performance comparison of different scalar functions used for 3D keypoint detection on the entire Bosphorus database

	Avg. number	HOG (%)	HOS (%)	HOGS (%)	HOMQ (%)
$C_H(\mathbf{p})$	322	76.2	83.6	74.6	88.5
$C_M(\mathbf{p})$	293	69.6	82.5	71.3	87.3
$C_m(\mathbf{p})$	355	76.2	85.5	75.5	91.0
$C_M(\mathbf{p}) + C_m(\mathbf{p})$	648	82.5	90.1	81.9	94.1

Table 4 Evaluation the detector fusion using the fine-grained matcher (FGM) and evaluated on the entire Bosphorus database

	Avg. number	HOG (%)	HOS (%)	HOGS (%)	HOMQ (%)
$C_M(\mathbf{p})$	293	86.7	91.8	91.9	94.8
$C_m(\mathbf{p})$	355	90.8	92.8	88.8	95.4
$C_M(\mathbf{p}) + C_m(\mathbf{p})$	648	93.1	94.4	93.1	96.6

Table 5 Performance comparison of different 3D keypoint descriptors: Spin Image (Johnson and Hebert 1999), 3D Tensor (Mian et al. 2008), meshSIFT (Smeets et al. 2013) and HOMQ (this paper) on the entire Bosphorus database

	Spin Image	3D Tensor	meshSIFT	HOMQ
$C_M(\mathbf{p})$	57.4 %	64.1 %	83.6 %	87.3 %
$C_m(\mathbf{p})$	58.0 %	69.0 %	86.7 %	91.0 %
$C_M(\mathbf{p}) + C_m(\mathbf{p})$	68.5 %	79.7 %	90.2 %	94.1 %

tions in 3D keypoint detection, and their accuracies are compared on the whole Bosphorous database. For a fair comparison, their corresponding descriptors are matched by the same CGM. From Table 3, we can find that the keypoint detector $C_m(\mathbf{p})$ using the minimum principal curvature locates more keypoints on average than $C_H(\mathbf{p})$ using the mean curvature which, in turn, beats $C_M(\mathbf{p})$ using the maximum principal curvature. Moreover, regardless of the local shape descriptor used, the 3D face recognition framework using $C_m(\mathbf{p})$ outperforms $C_H(\mathbf{p})$ which, in turn, beats $C_M(\mathbf{p})$. Finally, the score-level fusion of two principal curvature detectors, $C_M(\mathbf{p})$ and $C_m(\mathbf{p})$, significantly outperforms the results of the mean curvature detector, $C_H(\mathbf{p})$, while making use of 648 keypoints on average which doubles the averaged number, 322, located by the $C_H(\mathbf{p})$ detector. A possible reason is that when using the mean curvature as in meshSIFT, the keypoint detector fails to locate some *elliptic* and *hyperbolic* points with large principal curvatures but their mean curvatures are close to 0. It's worth noting that this effectiveness of the score-level fusion, $C_M(\mathbf{p}) + C_m(\mathbf{p})$, is also valid for FGM. For example, the performance of HOG descriptor is improved from 86.7 % for $C_M(\mathbf{p})$ and 90.8 % for $C_m(\mathbf{p})$, to 93.1 % for their fusion (see Table 4).

6.2 3D Keypoint Descriptors: Discriminative Power

As shown in Table 1, we can see that the HOMQ descriptor, which encodes the local shape properties through feature-level fusion of HOG, HOS, and HOGS, achieves the best matching performance and thus has the strongest discriminative power. Meanwhile, we also compare its discriminative power to other state-of-art 3D keypoint descriptors, including the spin image (Johnson and Hebert 1999), Mian's 3D tensor descriptor (Mian et al. 2008), as well as the meshSIFT descriptor (Smeets et al. 2013). To be fair, the comparison is conducted under a common framework. We apply the same 3D keypoint detector (i.e. two principal curvatures individually, and their score-level fusion), and the same 3D keypoint matcher (CGM) with the same parameters.

For the spin image, we directly use the code *calcSpinImages.m*[1].; for Mian's 3D tensor descriptor, we reproduce the method according to Mian et al. (2008); and for the meshSIFT descriptor, we adopt the code *meshSIFT*[2]. The comparative results are shown in Table 5. From this table, we can see that the proposed HOMQ descriptor always achieves the best results, with an improvement of about 4 % over the meshSIFT descriptor, which takes the second place. The spin image obtains the lowest performance. These results, once more, prove that the proposed HOMQ descriptor, which comprises multi-order surface differential quantities, has very strong discriminative power.

6.3 3D Keypoint Matchers: Similarity Measurements

To further analyze in depth the intuitive impression that FGM is better than CGM (i.e. the finer the similarity measurement, the stronger the matcher), we consider two groups of typical

[1] http://www.csse.uwa.edu.au/~ajmal/code.html.

[2] https://perswww.kuleuven.be/~u0059456/meshSIFT.html.

Table 6 Comparison of the normalized similarity measurements between CGM and FGM

	CGM Probe A1	FGM Probe A1	CGM Probe A2	FGM Probe A2
Gallery A	0.11	0.13	0.35	0.64
Gallery B	0.14	0.05	0.08	0.04

Gallery A, Probes A1 and A2 are three different scans of the same person, and Gallery B belongs to another person

Table 7 Estimated computational costs in seconds for identification of a single probe using a gallery set of 105 subjects from the Bosphorus database

	Detection $C_M(\mathbf{p})(s)$ or $C_m(\mathbf{p})$ (s)	Description HOMQ (s)	Matching CGM vs. FGM (s)
Total time: $C_M(\mathbf{p})$	10.2	51.0	2.1 vs. 3.9
Total time: $C_m(\mathbf{p})$	10.2	59.5	2.5 vs. 4.6

matching examples as shown in Table 6. In the first group, CGM fails to recognize Probe A1, while FGM succeeds. In contrast, both CGM and FGM correctly recognize Probe A2 in the second group. Their normalized similarity measurements are reported in Table 6, based on the score-level fusion of two principal curvature detectors and the HOMQ descriptor. The normalized similarity measurements range from 0 to 1, and the larger the score, the more similar the faces. Note that the average reconstruction errors in FGM should be further subtracted by 1 after normalization. From Table 6, we can find out that, compared with CGM, FGM significantly enlarges intra-class similarity (Gallery A vs. Probe A2) and reduces inter-class similarity (Gallery B vs. Probe A1).

6.4 Generalization Ability

To evaluate the generalization ability of the proposed method, we test our method on the FRGC v2.0 database (Phillips et al. 2005), which contains 4007 nearly frontal 3D face scans of 466 subjects with different expressions: neutral, anger, happiness, surprise, sadness, disgust, and puffy faces. The face scans are captured by the Minolta Vivid 900 laser scanner under controlled lighting conditions. Compared with the Bosphorus database, the FRGC v2.0 is the first largest public 3D face database, and has been widely used for the assessment of 3D face recognition algorithms, especially with respect to facial expression variations. All the scans of FRGC v2.0 are first preprocessed using the *3D Face Preprocessing Tools* developed by Szeptycki et al. (2009). The preprocessing pipeline contains: spike and noise removing, hole filling, nose tip localization, face cropping and triangulation.

To ensure a fair comparison, we use the same experimental protocol as in Smeets et al. (2013) (i.e. first vs. all). By using the score-level fusion of two principal curvature detectors and the HOMQ descriptor, our algorithm achieves identification rates of 93.3 and 96.3 % for CGM and FGM, respectively. These results largely outperform the score of 89.6 % of the meshSIFT reported in Smeets et al. (2013). Note that both methods are registration-free and based on the SIFT-like framework of 3D keypoint detection, description and matching. These results demonstrated that the proposed method has a good ability of generalization in different databases.

6.5 Time Cost

The proposed method is currently developed using MATLB on the Windows 7 platform, and all the experiments are implemented on a PC with the CPU by Intel 950, 3.07 GHz, 8GB RAM. Table 7 lists the estimated computational costs of individual steps of our method for identifying a single probe from a gallery set of 105 subjects from the Bosphorus database. Notice that the total expenditure for recognizing a probe face generally depends on the number of keypoints detected and used for the following description and matching. Here, for the detection module, we assume that for each probe, 300 and 350 keypoints are detected by $C_M(\mathbf{p})$ and $C_m(\mathbf{p})$, respectively (see Table 3).

From Table 7, we can find that the description module is the most time-consuming part, taking nearly 1 min (59.5 s) in the case when $C_m(\mathbf{p})$ is used. This is mainly due to the computation of neighboring points in a geodesic disc with a radius of 22.5 mm when building the HOMQ descriptor. The main time cost for the detection module is the computation of principal curvatures in different scale spaces, which takes about 10 s. For the matching module, CGM roughly runs two times faster than FGM. It is worth noting that the expenditures reported here for CGM and FGM are those used to match a probe face scan to all the 105 gallery face scans. In practice, parallel computation can be investigated for $C_M(\mathbf{p})$ and $C_m(\mathbf{p})$, and, in this case, our method can finish the task of identifying a single probe in around 1.25 mins.

7 Conclusion

We present a novel mesh version of the SIFT-like algorithm for registration-free 3D face recognition under expression changes, occlusions, and pose variations. We propose a principal curvature-based 3D keypoint detection algorithm, which can repeatedly identify complementary locations on a face scan where local curvatures are high. Moreover, a robust 3D local coordinate system is built at each keypoint allowing extraction of rotation and translation invariant 3D keypoint descriptors. Three local descriptors, corresponding to three different surface differential quantities, and their feature-level fusion, are proposed and compared. These single-quantity based descriptors contain strong complementarity information, enabling their fusion to provide a comprehensive and discriminative description of local shape. We also propose a multi-task SRC based fine-grained 3D keypoint matching algorithm and compare it with the SIFT-like coarse-grained matching scheme. Our algorithm is tested on the Bosphorus database and achieves identification rates of 96.56 % (entire database), 98.82 % (expression subset), 99.21 % (occlusion subset), and 91.14 % (pose subset), respectively. Meanwhile, good generalization ability is exhibited on the FRGC v2.0 database. Moreover, our algorithm has a potential for further improvements in speed, if upgraded by more efficient keypoint detectors (e.g., random sampling) and in also accuracy, if integrated with more effective keypoint descriptors and/or matchers. In In addition, we will evaluate the proposed approach for general 3D object description, matching and recognition.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 11401464, 61202237 and 61273263; the China Postdoctoral Science Foundation (No. 2014M560785); the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20121102120016); the French research agency, Agence Nationale de la Recherche (ANR) under Grant ANR-07-SESU-004, ANR-2010-INTB-0301-01 and ANR-13-INSE-0004-02; the joint project by the LIA 2MCSI lab between the group of Ecoles Centrales and Beihang University; and the Fundamental Research Funds for the Central Universities. We would like to thank the Bosphorus (Savran et al. 2008) and the FRGC (Phillips et al. 2005) organizers for the face data, Peyré for the *Toolbox Fast Marching*.

References

Al-Osaimi, F., Bennamoun, M., & Mian, A. (2009). An expression deformation approach to non-rigid 3d face recognition. *International Journal of Computer Vision*, *81*(3), 302–316.

Alyüz, N., Gökberk. B., & Akarun. L. (2008). A 3d face recognition system for expression and occlusion invariance. In: *2nd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)* (pp 1–7).

Alyüz, N., Gökberk, B., & Akarun, L. (2010). Regional registration for expression resistant 3-d face recognition. *IEEE Transactions on Information Forensics and Security*, *5*(3), 425–440.

Alyüz, N., Gokberk, B., & Akarun, L. (2013). 3-d face recognition under occlusion using masked projection. *IEEE Transactions on Information Forensics and Security*, *8*(5), 789–802.

Berretti, S., Werghi, N., del Bimbo, A., & Pala, P. (2013). Matching 3d face scans using interest points and local histogram descriptors. *Computers and Graphics*, *37*(5), 509–525.

Bowyer, K. W., Chang, K., & Flynn, P. (2006). A survey of approaches and challenges in 3d and multi-modal 3d+2d face recognition. *Computer Vision and Image Understanding*, *101*, 1–15.

Colombo, A., Cusano, C., & Schettini, R. (2011). Three-dimensional occlusion detection and restoration of partially occluded faces. *Journal of Mathematical Imaging and Vision*, *40*(1), 105–119.

Drira, H., Ben Amor, B., Srivastava, A., Daoudi, M., & Slama, R. (2013). 3d face recognition under expressions, occlusions, and pose variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(9), 2270–2283.

Faltemier, T. C., Bowyer, K. W., & Flynn, P. J. (2008). A region ensemble for 3d face recognition. *IEEE Transactions on Information Forensics and Security*, *3*(1), 62–73.

Goldfeather, J., & Interrante, V. (2004). A novel cubic-order algorithm for approximating principal direction vectors. *ACM Transactions Graphics*, *23*(1), 45–63.

Guo, Z., Zhang, Y., Xia, Y., Lin, Z., Fan, Y., & Feng, D. (2013). Multi-pose 3d face recognition based on 2d sparse representation. *Journal of Visual Communication and Image Representation*, *24*(2), 1047–3203.

Huang, D., Ardabilian, M., Wang, Y., & Chen, L. (2012). 3-d face recognition using elbp-based facial description and local feature hybrid matching. *IEEE Transactions on Information Forensics and Security*, *7*(5), 1551–1565.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cats visual cortex. *The Journal of Physiology*, *160*, 106–154.

Johnson, A., & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(5), 433–449.

Kakadiaris, I. A., Passalis, G., Toderici, G., Murtuza, M. N., Lu, Y., Karampatziakis, N., et al. (2007). Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(4), 640–649.

Li, H., Huang, D., Lemaire, P., Morvan, J., & Chen, L. (2011). Expression robust 3d face recognition via mesh-based histograms of multiple order surface differential quantities. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)* (pp. 4019–4023).

Li, H., Huang, D., Morvan, J. M., Chen, L., & Wang, Y. (2014). Expression-robust 3d face recognition via weighted sparse representation of multi-scale and multi-component local normal patterns. *Neurocomputing*, *133*, 179–193.

Li, S. Z., & Jain, A. K. (2005). *Handbook of Face Recognition*. Secaucus, NJ: Springer-Verlag New York Inc.

Li, X., Jia, T., & Zhang, H. (2009). Expression-insensitive 3d face recognition using sparse representation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2575–2582).

Liao, S., Jain, A., & Li, S. (2013). Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(5), 1193–1205.

Lo, T., & Siebert, J. (2009). Local feature extraction and matching on range images: 2.5d sift. *Computer Vision and Image Understanding*, *113*(12), 1235–1250.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Maes, C., Fabry, T., Keustermans, J., Smeets, D., Suetens, P., & Vandermeulen, D. (2010). Feature detection on 3d face surfaces for pose

normalisation and recognition. In: Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), pp 1–6.

Masi, I., Lisanti, G., Bagdanov, A., Pala, P., & Del Bimbo, A. (2013). Using 3d models to recognize 2d faces in the wild. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 775–780).

Meyer, T., Eriksson, M., & Maggio, R. (2001). Gradient estimation from irregularly spaced data sets. *Mathematical Geology*, *33*, 693–717.

Mian, A. S., Bennamoun, M., & Owens, R. A. (2007). An efficient multimodal 2d–3d hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(11), 1927–1943.

Mian, A. S., Bennamoun, M., & Owens, R. A. (2008). Keypoint detection and local feature matching for textured 3d face recognition. *International Journal of Computer Vision*, *79*(1), 1–12.

Mohammadzade, H., & Hatzinakos, D. (2013). Iterative closest normal point for 3d face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(2), 381–397.

Ocegueda, O., Passalis, G., Theoharis, T., Shah, S., & Kakadiaris, I. (2011). Ur3d-c: Linear dimensionality reduction for efficient 3d face recognition. In: *Proceedings of IEEE International Joint Conference on Biometrics (IJCB)*.

Passalis, G., Perakis, P., Theoharis, T., & Kakadiaris, I. (2011). Using facial symmetry to handle pose variations in real-world 3d face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(10), 1938–1951.

Pati, Y.C., Rezaiifar, R., & Krishnaprasad, P.S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers (ACSSC)*.

Phillips, P., Flynn, P. J., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., & Worek, W. (2005). Overview of the face recognition grand challenge. In: *Proeedings of of IEEE conference of Computer Vision and Pattern Recognition (CVPR)*.

Queirolo, C., Silva, L., Bellon, O., & Segundo, M. (2010). 3d face recognition using simulated annealing and the surface interpenetration measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(2), 206–219.

Savran, A., Alyüz, N., Dibeklioglu, H., Çeliktutan, O., Gökberk, B., Sankur, B., & Akarun, L. (2008). 3d face recognition benchmarks on the bosphorus database with focus on facial expressions. In: *Proceedings of Workshop on Biometrics and Identity Management (WBIM)*.

Skellya, L. J., & Sclaroffb, S. (2007). Improved feature descriptors for 3-d surface matching. *Proceedings of of SPIE Two- and Three-Dimensional Methods for Inspection and Metrology*, *6762*, 1–12.

Smeets, D., Claes, P., Hermans, J., Vandermeulen, D., & Suetens, P. (2012). A comparative study of 3-d face recognition under expression variations. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *42*(5), 710–727.

Smeets, D., Keustermans, J., Vandermeulen, D., & Suetens, P. (2013). meshsift: Local surface features for 3d face recognition under expression variations and partial data. *Computer Vision and Image Understanding*, *117*(2), 158–169.

Spreeuwers, L. (2011). Fast and accurate 3d face recognition using registration to an intrinsic coordinate system and fusion of multiple region classifiers. *International Journal of Computer Vision*, *93*(3), 389–414.

Szeptycki, P., Ardabilian, M., & Chen, L. (2009). A coarse-to-fine curvature analysis-based rotation invariant 3d face landmarking. In: *International Conference on Biometrics: Theory, Applications and Systems (ICB)* (pp. 3206–3211).

Tola, E., Lepetit, V., & Fua, P. (2010). Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(5), 815–830.

Tombari, F., Salti, S., & Di Stefano, L. (2013). Performance evaluation of 3d keypoint detectors. *International Journal of Computer Vision*, *102*(1–3), 198–220.

Veltkamp, R.C., van Jole, S., Drira, H., Amor, B.B., Daoudi, M., Li, H., Chen, L., Claes, P., Smeets, D., Hermans, J., Vandermeulen, D., & Suetens, P. (2011). Shrec '11 track: 3d face models retrieval. In: *Euro-graphics Workshop on 3D Object Retrieval (3DOR)* (pp. 89–95).

Wang, Y., Liu, J., & Tang, X. (2010). Robust 3d face recognition by local shape difference boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(10), 1858–1870.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 210–227.

Yang, A.Y., Wright, J., Ma, Y., & Sastry, S.S. (2007). Feature selection in face recognition: A sparse representation perspective. Tech. Rep. UCB/EECS-2007-99, EECS Department, University of California, Berkeley. http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-99.html.

Zaharescu, A., Boyer, E., Varanasi, K., & Horaud, R. (2009). Surface feature detection and description with applications to mesh matching. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 373–380).

Zaharescu, A., Boyer, E., & Horaud, R. P. (2012). Keypoints and local descriptors of scalar functions on 2d manifolds. *International Journal of Computer Vision*, *100*(1), 78–98.

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys*, *35*, 399–458.