3D facial expression recognition using SIFT descriptors of automatically detected keypoints

Stefano Berretti, Boulbaba Ben Amor, Mohamed Daoudi, Alberto del Bimbo

Vis Comput (2011) 27:1021–1036
DOI 10.1007/s00371-011-0611-x

ORIGINAL ARTICLE

3D facial expression recognition using SIFT descriptors of automatically detected keypoints

Stefano Berretti · Boulbaba Ben Amor · Mohamed Daoudi · Alberto del Bimbo

Published online: 30 June 2011

Abstract Methods to recognize humans' facial expressions have been proposed mainly focusing on 2D still images and videos. In this paper, the problem of person-independent facial expression recognition is addressed using the 3D geometry information extracted from the 3D shape of the face. To this end, a completely automatic approach is proposed that relies on identifying a set of facial keypoints, computing SIFT feature descriptors of depth images of the face around sample points defined starting from the facial keypoints, and selecting the subset of features with maximum relevance. Training a Support Vector Machine (SVM) for each facial expression to be recognized, and combining them to form a multi-class classifier, an average recognition rate of 78.43% on the BU-3DFE database has been obtained. Comparison with competitor approaches using a common experimental setting on the BU-3DFE database shows that our solution is capable of obtaining state of the art results. The same 3D face representation framework and testing database have been also used to perform 3D facial expression retrieval (i.e., retrieve 3D scans with the same facial expression as shown by a target subject), with results proving the viability of the proposed solution.

Keywords 3D facial expression recognition · 3D facial expression retrieval · SIFT keypoints · Feature selection · SVM classification

S. Berretti (✉) · A. del Bimbo
Dipartimento di Sistemi e Informatica, University of Firenze, Firenze, Italy
e-mail: stefano.berretti@unifi.it

A. del Bimbo
e-mail: alberto.delbimbo@unifi.it

B. Ben Amor · M. Daoudi
Institut TELECOM, TELECOM Lille 1, LIFL (UMR 8022), Lille, France

B. Ben Amor
e-mail: boulbaba.benamor@telecom-lille1.eu

M. Daoudi
e-mail: mohamed.daoudi@telecom-lille1.eu

1 Introduction

In recent years, automatic recognition of facial expressions has been an active research field targeting applications in several different areas such as Human–Machine Interaction, Computer Graphics and Psychology. The first studies on this subject date back to the late 1970s with the pioneering work of Ekman [6]. In these studies, it is evidenced that the *basic* facial expressions can be categorized into six classes, namely, *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*, plus the *neutral* expression. This expressions categorization has been also proved to be consistent across different ethnic groups and cultures.

The *Facial Action Coding System* was developed by Ekman and Friesen [7] to code the facial expressions through the movement of face points as described by the *action units*. This work inspired many researchers to analyze facial expressions in 2D by tracking facial features and measuring the amount of facial movements in still images and videos. Almost all of the methods developed in 2D use distributions of facial features as inputs to classification systems, and the outcome is one of the facial expression classes. These approaches mainly differ in the facial features selected and the classifier used to distinguish among the different facial expressions.

Recently, owing to the increasing availability of effective devices capable of acquiring high-resolution 3D data,

there has been a progressive shift from 2D to 3D approaches in order to perform face recognition and facial expression recognition. The main motivation is the robustness of 3D facial shape to illumination changes, pose and scale variations. Although many solutions have appeared to perform 3D face recognition (see [2, 11, 12, 18, 24, 27]), still few works have taken advantage of the 3D facial geometric information to perform facial expression recognition. The initial solutions to automatically perform facial expression recognition based on 3D face scans used very small databases and categorized just a few facial expressions [25]. Recently, the availability of new facial expression databases, like those constructed at the *Binghamton University* (BU-3DFE database) [34], and at the *Boğaziçi University* (Bosphorus database) [28], has pushed the research on this topic. In particular, the BU-3DFE database has become the de facto standard for comparing facial expression recognition algorithms. This is due to the fact that, differently from other 3D face data sets, the BU-3DFE database provides a precise categorization of facial scans according to Ekman's six basic facial expressions plus the neutral one, also providing different levels of expression intensities.

1.1 Previous work

Most of the works on 3D facial expression recognition can be categorized as based on *generic facial model* or *feature classification*.

In the first category, a general 3D face model is trained with prior knowledge, such as feature points, shape and texture variations, or local geometry labels. A dense correspondence between 3D faces is usually required to build the generic model. For example, in [25] a correspondence is established between faces with expression and their neutral pair by minimizing an energy function. A *Morphable Expression Model* (MEM) is constructed by applying *Principal Component Analysis* (PCA) to different expressions, so that new expressions can be projected into points in a low-dimensional space constructed by the eigen-expressions obtained by MEM. Expression classification is performed by comparing the *Euclidean* distances among projected points in the eigen-expression space, and a recognition rate of over 97% is reported on a small and private data set (just 25 subjects with 4 expressions per subject are included in the data set). An approach inspired by the advances of *Ant Colony Optimization* (ACO) and *Particle Swarm Optimization* (PSO) is proposed in [19]. In this work, first anatomical correspondence between faces is established using a generic 3D deformable model and the 83 manually detected facial landmarks of the BU-3DFE database. Then, surface points are used as a basis for classification, according to a set of classification rules which are discovered by an ACO/PSO based rule discovery algorithm. The performance of the algorithm evaluated on the BU-3DFE database scored a total recognition rate of 92.3%. In [20], face recognition and facial expression recognition are performed jointly by decoupling identity and expression components with a bilinear model. An elastically deformable model algorithm that establishes correspondence among a set of faces is proposed. Construction of the model relies on manually identified landmarks which are used to establish points correspondence in the training stage. Fitting these models to unknown faces enables face recognition invariant to facial expressions and facial expression recognition with unknown identity. A quantitative evaluation of the technique is conducted on the BU-3DFE database with an overall 90.5% facial expression recognition. In [10], the shape of an expressional 3D face is approximated as the sum of a basic facial shape component, representing the basic face structure and neutral-style shape, and an expressional shape component that contains shape changes caused by facial expressions. The two components are separated by first learning a reference face for each input non-neutral 3D face, and then, based on the reference face and the original expressional face, a facial expression descriptor is constructed which accounts for the depth changes of rectangular regions around eyes and mouth. Average recognition rates of 71.63% and 76.22% have been reported on the BU-3DFE database, respectively, by not using and using a reference neutral scan for each subject.

Approaches in the second category extract features from 3D face scans and classify them into different expressions. In [33], a feature-based facial expression descriptor is proposed and the BU-3DFE database is used for the first time. The face is subdivided into 7 regions using manually annotated landmarks, and primitive surface features are classified into basic categories such as *ridge*, *ravine*, *peak*, *saddle*, etc., using surface curvatures and their principal directions. The authors reported the highest average recognition rate of 83.6% using the primitive facial surface features and a LDA classifier. The facial expressions of *happiness* and *surprise* were reported to be the best identified with accuracies of 95% and 90.8%, respectively. Comparison with the results obtained using the *Gabor-wavelet* and the *Topographic Context* 2D appearance feature-based methods on the same database showed that the 3D solution outperforms the 2D methods. 3D facial expression recognition on the BU-3DFE database has been also performed in [29]. Among the 83 facial landmarks labeling the 3D faces of the BU-3DFE database, only six distance measures maximizing the differences of facial expressions are selected. These six distance values are used to form a distance vector for the representation of facial expressions as defined by the MPEG-4 *Facial Definition Parameter Set* [22]. The results obtained from a neural network classifier using the 3D distance vectors reach up to 98.3% in the recognition of *surprise* facial

expression, whereas the average recognition performance is 91.3%. In [30], a set of candidate features composed of normalized *Euclidean* distances between the 83 facial landmarks of the BU-3DFE database are first extracted. Then, a feature selection method based on maximizing the average relative entropy of marginalized class-conditional feature distributions is used to retain just the most informative distances. Using a regularized multi-class *AdaBoost* classification algorithm, a 95.1% average recognition rate for the six basic facial expressions is obtained on a subset of the BU-3DFE database. The neutral facial expression is not classified; rather, as a preprocessing step, its features serve as fiducial measures that are subtracted from the features of the six basic facial expressions of the corresponding subject. The approach proposed in [32] uses a modified PCA to classify facial expressions using only the shape information at a finite set of fiducial points which are extracted from the 3D neutral and expressive faces of the BU-3DFE database. The approach uses 2D texture images of the face to mark interest regions around the eyebrows, eyes, nose and mouth, and extracts facial contours in those regions with the help of an active contour algorithm. Then, these contours are uniformly sampled and the sampled points are mapped onto the 3D data set in order to generate a shape and color descriptor of the interest-regions. An average recognition rate of 81.67% is reported. In [14, 15] an approach based on the shape analysis of local facial patches is proposed. The patches are extracted around the 83 manually annotated facial landmarks of the BU-3DFE database, and the shape of each patch is described by a set of curves representing the surface points at the same Euclidean distance from the landmark. A Riemannian framework is then applied to compare the shape of curves undergoing different facial expressions. The length of the geodesic path that separates corresponding curves provides a quantitative information about their shape similarity. The best expression recognition results in the BU-3DFE database have been obtained using these measures as entries of a *Multiboost* classifier.

From the above review, it emerges that the large part of existing works on 3D facial expression recognition rely on the presence of landmarks accurately identified on the face. Methods based on *generic facial model* use landmarks to establish correspondences between faces in the construction of a deformable template face. Usually, these approaches are also computationally demanding due to the deformation process. Solutions based on *feature classification* in many cases compute distances between landmarks and evaluate how these distances change between expressional and neutral scans. The fact is that several landmarks are not automatically detectable and the precision required for their positioning demands for manual annotation in both training and testing stages. Furthermore, several solutions require a neutral scan for each subject in order to evaluate the differences generated in the 3D scans by facial expressions with respect to neutral reference scans. In practice, these factors limit the applicability of many approaches.

1.2 Contribution and paper organization

A few recent works have shown that local descriptors computed at salient keypoints can be usefully applied to describe 3D objects. In [18], a 3D keypoint detector and descriptor inspired by the *Scale Invariant Feature Transform* (SIFT) [13], has been designed and used to perform 3D face recognition through a hybrid 2D+3D approach that also uses the SIFT detector and descriptor to index 2D textured face images. In [16], SIFTs are used to detect and represent salient points in multiple 2D range images derived from 3D face models for the purpose of 3D face recognition. A similar idea is used in [21] to perform 3D object retrieval by visual similarity, but in this case points of a sampling grid are used, and SIFT descriptors are computed for them. Finally, SIFT descriptors have been also used in [35] to perform 2D expression recognition from non-frontal face images.

Based on these studies, in this work we propose to use local descriptors to perform person-independent 3D facial expression recognition. Differently from existing approaches, we define a completely automatic solution that first detects a set of facial keypoints, and then exploits the local characteristics of the face around a set of sample points automatically derived from the facial keypoints. In particular, SIFT descriptors are computed around the sample points of the face, are combined together and used as a feature vector to represent the face. Before performing classification of the extracted descriptors, a feature selection approach is used to identify a subset of features with *minimal-redundancy* and *maximal-relevance* among the large set of features extracted with SIFT. The set of selected features is finally used to feed a set of classifiers based on *Support Vector Machines* (SVM). As it emerges from the experimental evaluation, the proposed approach is capable of achieving state of the art results on the BU-3DFE database just relying on few keypoints that are automatically detected and without using neutral scans as reference. In addition, we used the proposed face representation framework to perform experiments of 3D facial expression retrieval. The idea is to use the 3D scan of a subject with a target facial expression as query, and retrieve all the 3D scans of subjects that show the same facial expression as the target one. This retrieval scenario can have practical applications in several different contexts, such as in Facial Character Animation, Psychology studies, Medical Aesthetic, and so on. However, we did not find any retrieval experiment in previous works on the analysis of 3D facial expressions; results of our approach show its viability also in this task.

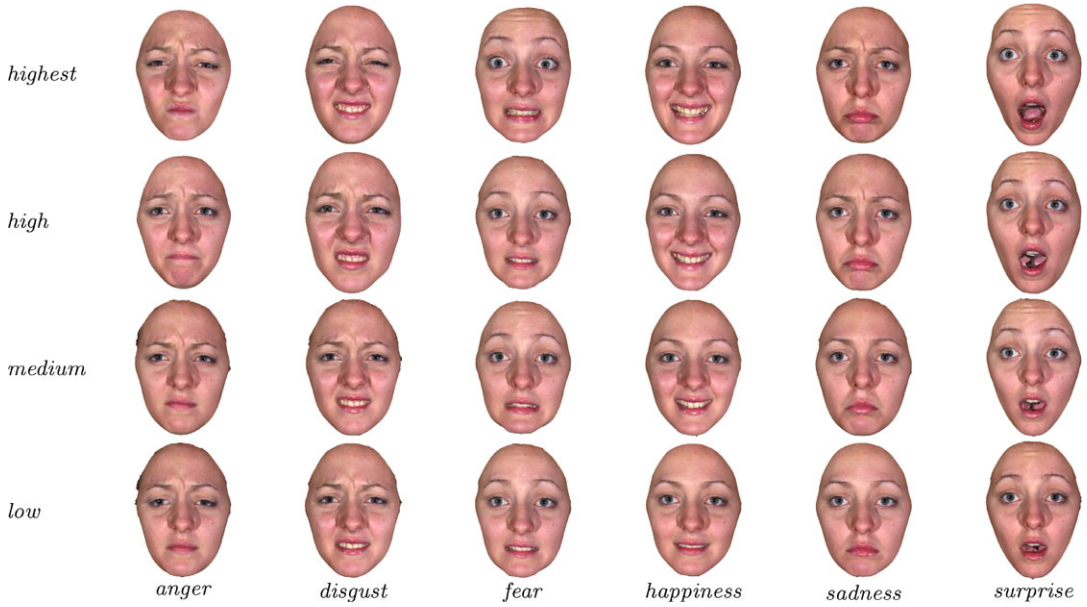


Fig. 1 BU-3DFE database: 3D face scans (with texture) of a sample subject showing the six basic facial expressions at the four levels of intensity (from *highest* to *low*)

This work develops on our preliminary results presented in [2], by extending the approach to a completely automatic solution. To the best of our knowledge this is the first work that proposes a fully automatic approach for person-independent 3D facial expression recognition and also provides results for 3D facial expression retrieval.

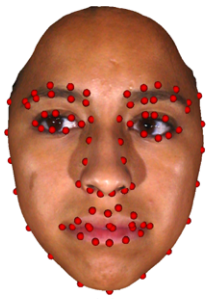
The rest of the paper is organized as follows: In Sect. 2, the characterizing elements of the BU-3DFE database are summarized in order to motivate some of the choices that guide our approach. A solution for the automatic identification of facial keypoints is presented in Sect. 3. In Sect. 4, the main characteristics of SIFT descriptors are described, and their adaptation to our case is presented. The feature selection approach used to reduce the set of SIFT features and the SVM-based classification of the selected features are addressed in Sect. 5. Experiments carried out with the proposed approach, with results and comparative evaluation are reported in Sect. 6. Finally, discussion and conclusions are given in Sect. 7.

2 The BU-3DFE database

The BU-3DFE database was recently constructed at *Binghamton University* [34]. It was designed to provide 3D facial scans of a large population of different subjects each showing a set of prototypical emotional states at various levels of intensities. 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* (NE) facial expression as well as the six basic facial expressions defined by Ekman, namely, *anger* (AN), *disgust* (DI), *fear* (FE), *happiness* (HA), *sadness* (SA), and *surprise* (SU). 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. 1 shows the six basic facial expressions of a sample 3D face at the four levels of intensity.

Each 3D facial expression scan is also associated with a raw 3D face mesh, a cropped 3D face mesh, a pair of texture images with two-angles of view (about $+45°$ and $-45°$ away from the face frontal normal), a frontal-view texture image, a set of 83 *manually annotated* facial landmarks, and a facial pose vector. These data give a complete 3D description of a face under a specific facial expression. The cropped and textured 3D face scan, and the 83 facial landmarks are shown in Fig. 2(a). It can be observed that the landmarks are distributed in correspondence with the most distinguishing traits of the face, i.e. *eyes*, *eyebrows*, *nose* and *mouth*, plus the face boundary, as summarized in Fig. 2(b). A more detailed description of the BU-3DFE database can be found in [34].

In this work, we only use the cropped 3D face scans in order to perform expression recognition. The 83 facial landmarks provided with each scan are considered just as ground-truth in order to validate the proposed solution for



(a)

	#points
right eye	8
left eye	8
right eyebrow	10
left eyebrow	10
nose	12
mouth	20
face boundary	15

(b)

Fig. 2 BU-3DFE database: (**a**) the 83 facial landmarks evidenced on a textured 3D face scan with neutral expression; (**b**) the table reports the number of manually identified landmarks for different regions of the face

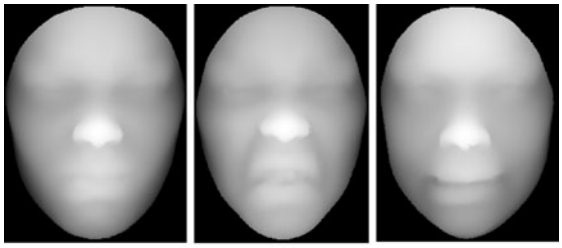
automatic identification of facial *keypoints* as discussed in the next section.

3 Automatic identification of facial keypoints

The BU-3DFE database is the standard benchmark to compare 3D facial expression recognition algorithms. However, the fact that this database provides a set of 83 manually identified landmarks on the 3D face scans, and the inherent difficulty in automatically detecting the majority of these landmarks both in 2D and 3D, has oriented the research towards semi-automatic solutions for 3D facial expression recognition. In these solutions, the position of facial landmarks is assumed to be known in order to achieve high facial expression recognition rates (see also the discussion on previous work in Sect. 1.1). In practice, this reduces the applicability of the existing solutions in the general case in which manual annotation of the landmarks in 3D is not available or even possible. To overcome this limitation, we propose a completely automatic solution to identify fiducial points of the face that in the following will be referred to as *keypoints*.

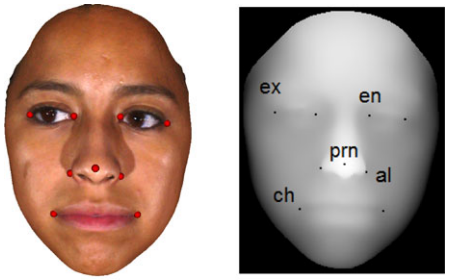
Some preprocessing was applied to the cropped 3D face scans before performing keypoint detection. First, spikes in the 3D face were removed using median filtering in the z-coordinate. Then, holes were filled using cubic interpolation and 3D scans were resampled on a uniform square grid at 0.7 mm resolution. The scans were also subjected to pose normalization by iteratively performing PCA alignment and resampling of the cropped portion of the face [17]. After these steps, the 3D face scans were transformed to *range images* where the gray value of each image pixel represents the depth of the corresponding point on the 3D surface. As an example, Fig. 3 shows the range images derived from the 3D face scans of a same subject under three different facial expressions.

On the range images, the point with maximum gray value has been used as initial estimate of the tip of the nose. This



(a) (b) (c)

Fig. 3 Range images derived from the 3D face scans of the same subject, for the expressions (highest level of intensity of the BU-3DFE database): (**a**) *anger*; (**b**) *disgust*; (**c**) *fear*



(a) (b)

Fig. 4 The 9 facial keypoints that can be automatically detected with our approach, shown on: (**a**) a textured 3D face scan; (**b**) the range image of the scan in (**a**)

point was used to crop a rectangular region of the face (following anthropometric statistical measures [8], the cropped region extends 50 mm on the left and 50 mm on the right of the nose tip, and 70 mm above and 50 mm below the nose tip). The cropped region of the face is used for all the subsequent steps of the processing.

Our approach starts from the consideration that just a few fiducial points of the face can be automatically identified with sufficient robustness across different individuals. This is supported by recent studies as that in [11], where methods are given to automatically identify 10 facial fiducial points on the 3D face scans of the *Texas 3D Face Recognition Database*. Following this idea, we propose a general method to automatically identify 9 keypoints of the face, namely, the tip of the nose (*pronasale, prn*), the two points that define the nose width (*alare, al*), the points at the *inner* and *outer* eyes (*endocanthion, en* and *exocanthion, ex*, respectively), and the outer mouth points (*cheilion, ch*), as evidenced on the 3D face scan and range image of Figs. 4(a) and (b).

We used different algorithms in order to detect the keypoints. For the nose tip and the two alare points we used solutions derived from the work in [11]. For the remaining points (inner and outer eyes and outer mouth) we used a solution based on the SIFT detector applied to local search windows.

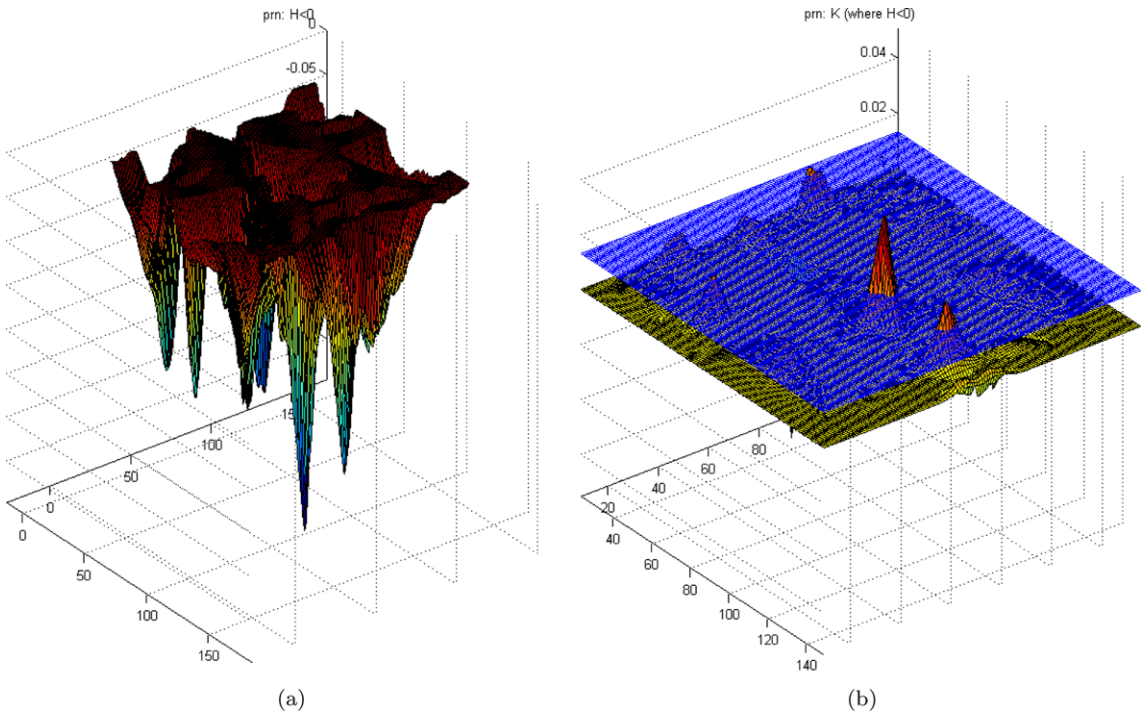


Fig. 5 The magnitude of the (**a**) mean curvature H, where $H < 0$; and of the (**b**) Gaussian curvature K where $H < 0$. A local maximum of the Gaussian curvature can be observed in correspondence with the nose tip. *The blue plane* shows a threshold equal to 1/2 of the maximum peak under which the local maxima are discarded

3.1 Nose tip and alare points

Given a 3D face scan in a frontal upright canonical pose, the following operations are applied:

Nose tip (*pronasale, prn*) The point of the range image with maximum gray value is used as initial estimate of the tip of the nose. This position is refined using the property of the surface curvature. The *Gaussian* surface curvature (K) and the *mean* surface curvature (H) of the facial range images are computed from their first and second derivatives [5]. In so doing, due to the sensitivity of second derivatives to surface noise, the surface is smoothed with a Gaussian filter and the surface is approximated using a biquadratic polynomial [4]. According to [11], the region surrounding the tip of the nose is *convex* ($H < 0$) and has high *elliptic* Gaussian curvature ($K > 0$) (see Figs. 5(a) and (b)). Following this observation, the nose tip is determined in the convex part of the central region of the face as the point with a local maximum of the elliptic Gaussian curvature which is closest to the initial estimate of the nose tip.

Nose width (*alare, al–al*) The search of the alare points is performed in a window of 50 mm width and 42 mm height centered on the nose tip [11]. In this region, the edges of the facial range images are identified using a *Laplacian of Gaussian* (LoG) edge detector with $\sigma = 3$ pixels. The edges of the left and right part boundaries of the nose are detected by traversing outwards horizontally in both directions from the tip of the nose and by retaining the first edges encountered. Then, the points along the nasal boundary with high negative curvature ("critical" points) are detected. In order to compute the boundary curvature, the contour is coded counterclockwise according to the *Freeman chain code* [26]. In order to eliminate the hard discontinuities due to the inversion of the phase of the chain code, the code is *unwrapped* [3]. Then, a *derivative of Gaussian* (doG) filter is applied to the chain code in order to smooth and derive it. This corresponds to good approximation to the curvature of the contour. Finally, the critical points are identified in correspondence with the local minima of the derivative that correspond to maxima of the curvature. The two alare points are selected as the outer left and outer right critical points. As an example, Fig. 6(a) shows the window centered on the nose tip used for the search of the alare points; Fig. 6(b) shows in blue the detected nose tip and the alare keypoints on the contour of the nose; in Fig. 6(c) the identification of the critical points using the doG of the unwrapped chain code is reported.

3.2 Inner and outer eyes, outer mouth

The detection of the remaining fiducial points (i.e., inner and outer eyes, outer mouth) proceeds in cascade using the location of the nose tip and the alare points to identify search

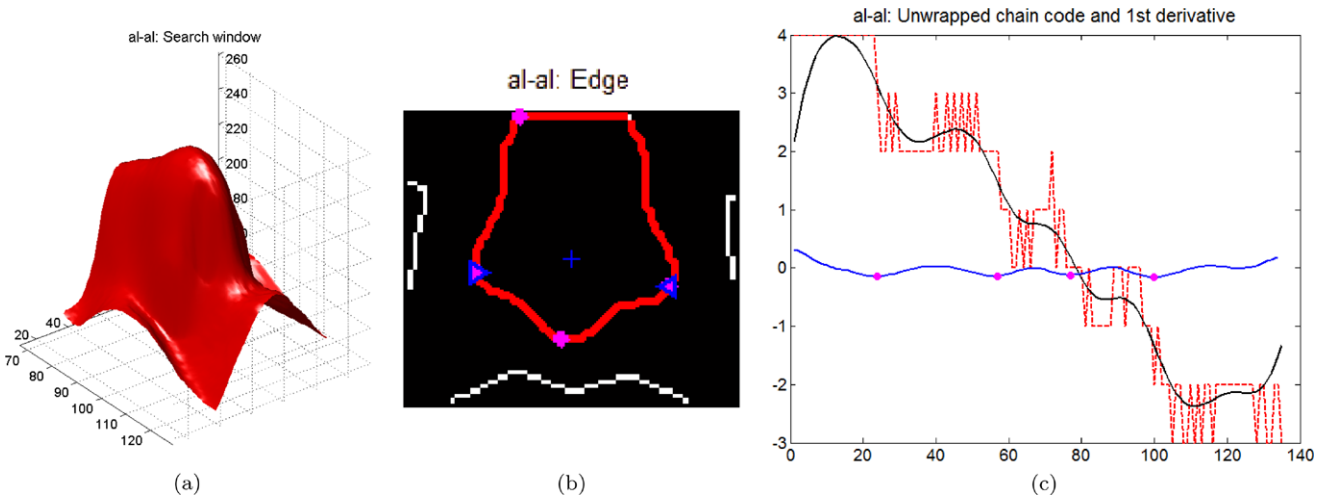


Fig. 6 (**a**) The window centered on the nose tip used for the search of the alare points; (**b**) the detected edges (*white line*), the contour of the nose (in *red*), the critical points (in *purple*), and the selected alare points (in *blue*) (the nose tip is also evidenced in *blue at the center of the window*); (**c**) the unwrapped chain code of the contour (*dashed red line*), the smoothed chain code (*black line*), its first derivative (*blue line*) and the critical points (in *purple*)

windows on the face. Following the studies of Farkas and Munro [9] and the application in [11], the search windows are determined as:

Inner eyes (*endocanthion, en–en*) The vertical extent (y-coordinate) of the search window in pixels is bounded by $0.126 \times |prn_y - v_y| \leq y - prn_y \leq 0.570 \times |prn_y - v_y|$, v being the upper point of the face. The horizontal bounds (x-coordinate) for the inner left and inner right eyes are $prn_x \leq x \leq al_x^{\text{left}} + 0.5 \times |al_x^{\text{left}} - al_x^{\text{right}}|$ and $al_x^{\text{right}} - 0.5 \times |al_x^{\text{left}} - al_x^{\text{right}}| \leq x \leq prn_x$, respectively.

Outer eyes (*exocanthion, ex–ex*) The initial positions of the outer corners of the left and right eyes are given by $(en_x^{\text{left}} + |en_x^{\text{left}} - en_x^{\text{right}}|, (en_y^{\text{left}} + en_y^{\text{right}})/2)$ and $(en_x^{\text{right}} - |en_x^{\text{left}} - en_x^{\text{right}}|, (en_y^{\text{left}} + en_y^{\text{right}})/2)$, respectively. A rectangular search window of 34 mm width and 20 mm height is then identified around these points.

Outer mouth (*cheilion, ch–ch*) The vertical limits of the mouth are determined by detecting the upper and lower lip regions with elliptic Gaussian curvature [11]. These regions are used to define the vertical limits of the search windows. The locations of the alare points are used to constrain horizontally the search window for the left and right mouth corners with the respective bounds, $al_x^{\text{left}} \leq ch_x^{\text{left}} \leq al_x^{\text{left}} + 0.7 \times |al_x^{\text{left}} - al_x^{\text{right}}|$ and $al_x^{\text{right}} - 0.7 \times |al_x^{\text{left}} - al_x^{\text{right}}| \leq ch_x^{\text{right}} \leq al_x^{\text{right}}$.

In [11], both the 2D texture and the 3D geometry of the face are used to identify the keypoints in these search windows. In our approach, we designed a completely 3D solution that only relies on 3D data to identify the keypoints. To this end, the SIFT *detector* algorithm is run in the *en–en*, *ex–ex* and *ch–ch* search windows. In fact, SIFT has been defined on 2D gray-scale images and includes a *keypoints detector* and a *feature extractor* [13]. By definition, keypoints detected by SIFT are mainly located at corner points of an image, so that they can be useful to capture significant anthropometric face points. The SIFT point detected at the *highest scale* is retained as keypoint of the search window.

As an example, Fig. 7(a) shows, for a sample subject, the keypoints detected with our approach and the search windows of the left part of the face for *en*, *ex* and *ch*. The 3D surface of these search windows and the keypoints identified on the corresponding range images are also shown in (b)–(d) of Fig. 7 for *en*, *ex* and *ch*, respectively. Each detected keypoint is represented by a circle of radius proportional to the scale at which the keypoint is detected (the orientation of the radius also shows the dominant orientation of the SIFT descriptor). It can be observed that a few keypoints are detected in each search window and the keypoints identified at the largest scale (i.e., centers of the largest circles in the range images on the right of Figs. 7(b) and (d)) well identify the searched fiducial points.

Experiments on the accuracy of keypoint detection are reported in Sect. 6.4.

4 SIFT features of facial sample points

The nine automatically detected keypoints are used as reference to derive a set of sampling points on the face. This is obtained by considering 8 lines that connect pairs of keypoints, as shown in Fig. 8(a). In particular, these lines con-

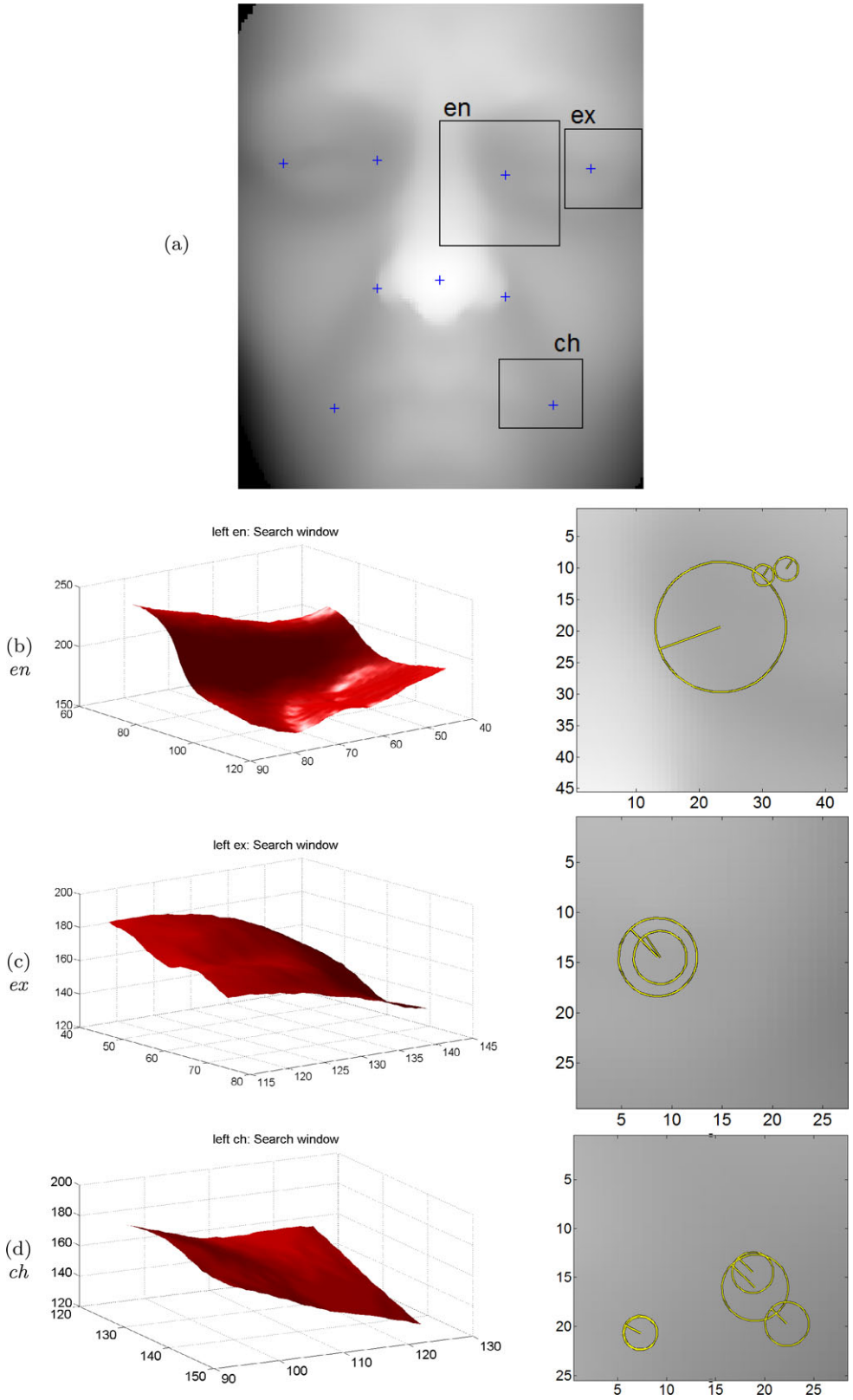
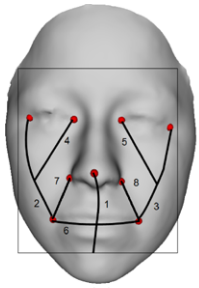


Fig. 7 (**a**) Keypoints detected and the search windows for *en*, *ex* and *ch in the left part of the face*. In (**b**), (**c**) and (**d**) the 3D rendering of the search windows in (**a**) are reported *on the left*, and the corresponding keypoints identified on the range images of the windows are reported *on the right*. In the range images *on the right of* (**b**), (**c**) and (**d**) a circle is centered on each SIFT detected keypoint in the window: the radius of the circle is proportional to the scale at which the keypoint is detected, whereas the orientation of the radius shows the dominant orientation of the SIFT descriptor. The keypoint with maximum radius (i.e., the keypoint detected at the largest scale) is retained as fiducial point in the window

nect the *nose tip* to the lower point of the face (*line 1*), the *outer mouth* with the *outer eyes* (*lines 2, 3*), the *inner eyes* with the mid points of lines 2 and 3 (*lines 4, 5*), the *outer mouth* points each other (*line 6*), and the *alare* points with the *outer mouth* (*lines 7, 8*). Lines are sampled uniformly with a different number of points as reported in the Table of



(a)

line	*#sample points*	*indices*
1	12	1 - 12
2	16	13 - 28
3	16	29 - 44
4	9	45 - 53
5	9	54 - 62
6	9	63 - 71
7	4	72 - 75
8	4	76 - 79

(b)

Fig. 8 (**a**) The eight lines along which the sample points are located (the cropped region of the face is also reported); (**b**) the number of sample points and their indices grouped according to the surface line they belong to

Fig. 8(b). According to this, the face is sampled with a total number of 79 points.

In order to capture salient features that characterize different facial expressions in 3D, we compute local descriptors around the 79 sample points of the face. The SIFT *feature extraction* algorithm has been used for this purpose, so as to derive SIFT *descriptors*. A detailed explanation of the SIFT keypoint detection and feature extraction is given in [13]. In the following, we just summarize the main idea of SIFT descriptor and its adaptation to our context. Briefly, a SIFT descriptor of a small image patch, for example of size 4×4, is computed from the gradient vector histograms of the pixels in the patch. There are 8 possible gradient directions, and therefore the total size of the SIFT descriptor is $4 \times 4 \times 8 = 128$ elements. This descriptor is normalized to enhance invariance to changes in illumination (not relevant in the case of range images), and transformed in other ways to ensure invariance to scale and rotation as well. These properties make the SIFT descriptor capable of providing a compact and powerful local representation of the range image and, as a consequence, of the face surface.

The following settings have been used for the extraction of SIFT descriptors (we employed the publicly available implementation of SIFT given in [31]):

- For each range image, SIFT descriptors are computed at the 79 sample points;
- At these points, SIFT descriptors are computed at scale equal to 3. In order to achieve invariance to image rotation, the descriptor is computed relative to an orientation given by the dominant direction of local image gradient evaluated at the assigned scale;
- The orientation histograms of 4×4 neighbor regions of each sample point are used to calculate the SIFT descriptor. By computing the 128-dimensional SIFT descriptor at each of the 79 sample points, a feature vector with 10112 components is obtained to represent each range image.

To reduce the dimensionality and improve the significance of the features, only the features with *maximal-relevance* and *minimal-redundancy* have been selected using the feature selection analysis reported in Sect. 5.

5 Selection of relevant SIFT features

Feature selection is mainly motivated by the *curse of dimensionality*, which states that in presence of a limited number of training samples, each one represented as a feature vector in R^n, the mean accuracy does not always increase with vector dimension (n). Rather, the classification accuracy increases until a certain dimension of the feature vector and then decreases. In other words, the higher the dimensionality of the feature space, the higher the number of training samples required to achieve the same classification accuracy. Therefore, the challenge is to identify m out of the n features which will yield similar, if not better, accuracies as compared to the case in which all the n features are used in a classification task.

In the proposed analysis, feature selection is performed using the *minimal-redundancy maximal-relevance* (mRMR) model [23]. For a given classification task, the aim of mRMR is to select a subset of features by taking into account the ability of features to identify the classification label, as well as the redundancy among the features. These concepts are defined in terms of the *mutual information* between features.

Given two discrete random variables x and y, taking values in $\{s_i\}_{i=1}^N$, their joint probability $P(x, y)$ and the respective marginal probabilities $P(x)$ and $P(y)$, the mutual information between x and y is defined as the difference between the *Shannon*'s entropy of x and the conditional entropy of x given y, that is: $I(x, y) = H(x) - H(x|y)$, where the entropy is used as measure of the uncertainty of a random variable. In practice, this expression states that if from the uncertainty of x is subtracted the uncertainty of x once y is known, the information (in bits) that the variable y provides about x is obtained. According to this, mutual information provides a measure of the dependency of variables, and can also be computed as:

$$I(x, y) = \sum_{i=1}^{N} \sum_{j=1}^{N} P(s_i, s_j) \log \frac{P(s_i, s_j)}{P(s_i) P(s_j)}. \tag{1}$$

The work in [23] proposes to jointly maximize the dependency between a feature variable x_i and the classification variable l and minimize the dependency between pairs of feature variables x_i, x_j. Thus, the task of feature selection is posed as selecting from the complete set of n features S_n, a subset S_m of $m < n$ features that maximizes:

$$\frac{1}{m} \sum_{x_i \in S_m} I(x_i, l) - \frac{1}{\binom{m}{2}} \sum_{x_i, x_j \in S_m} I(x_i, x_j). \tag{2}$$

This expression takes into account the relevance of features with the class label while penalizing redundancy among them. Since the search space of subsets of m elements in R^m is too big to be explored in practice, S_m is determined incrementally by means of a forward search algorithm. Having a subset S_{m-1} of $m-1$ features, the feature $x_i \in \{S_n - S_{m-1}\}$ that determines a subset $\{x_i, S_{m-1}\}$ maximizing (2) is added. It can be shown that this nested subset strategy is equivalent to iteratively optimizing the following condition:

$$\max_{x_i \in S_n - S_{m-1}} \left(I(x_i, l) - \frac{1}{m-1} \sum_{x_j \in S_{m-1}} I(x_j, x_i) \right). \tag{3}$$

Experiments in [23] show that for subsets of more than 20 features, the S_m obtained by this method achieves a more accurate classification performance than the subset obtained by maximizing the $I(S_m, l)$ value (that is, the mutual information between the whole subset of variables and the classification label l), while the required computation cost is significantly lower.

5.1 SVM classification

In our approach, the mRMR algorithm is applied to the set of 10112-dimensional feature vectors representing the faces. Each vector $v_f = (f_1, \ldots, f_{10112})$ is constructed by concatenating the 128-dimensional SIFT descriptors computed at the face sample points, orderly from 1 to 79. A data discretization is applied to the vectors as preprocessing step. This is obtained by computing the mean value μ_k and the standard deviation σ_k for every feature f_k. Then, discretized values $\hat{f}_k$ are obtained according to the following rule:

$$\hat{f}_k = \begin{cases} 2 & \text{if } f_k < \mu_k - \alpha \cdot \sigma_k, \\ 3 & \text{if } \mu_k - \alpha \cdot \sigma_k \le f_k \le \mu_k + \alpha \cdot \sigma_k, \\ 4 & \text{if } f_k > \mu_k + \alpha \cdot \sigma_k, \end{cases} \tag{4}$$

α being a parameter that regulates the width of the discretization interval (it is equal to 0.2 in our experiments). The overall set of discretized feature vectors is used to feed the mRMR algorithm so as to determine the features which are more relevant in discriminating between different facial expressions of 3D face scans of different subjects.

The facial expression recognition problem is a multi-classification task that, in our approach, is faced as a combination of separated instances of *one*-vs.-*all* classification subproblems. For each subproblem, face scans showing one expression are assumed as targets (positive examples), whereas all the other scans with any different expression are considered as negative examples. Repeatedly, the target expression is changed among the six basic expressions provided by the BU-3DFE database, so that the sets of positive and negative examples change. Due to this, mRMR feature selection is performed independently for each classification subproblem. In general, this results into different features providing the minimal-redundancy and maximal-relevance for the purpose of discriminating across different facial expressions. Then, just the most relevant features identified for every expression are retained from the original feature vectors in order to train the classifiers. This results into vectors $v_{\hat{f}}^{\text{expr}} = (\hat{f}_{p_1}, \ldots, \hat{f}_{p_{N\text{expr}}})$, where $p_1, \ldots, p_{N\text{expr}}$ are the indices of the feature components selected in the original vector, and the subscript is the label of a particular expression.

The selected features are then used to perform facial expression recognition using a *maxima rule* between six *one*-vs.-*all* SVM classifiers, each with a radial basis function kernel of standard deviation equal to one (the publicly available SVMLight implementation of SVM has been used: http://svmlight.joachims.org/).

6 Experimental results

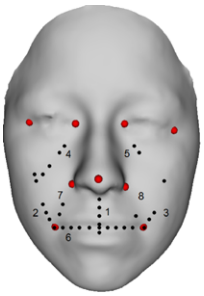
Experiments on the BU-3DFE database have been conducted using a setup similar to that in [10]. In particular, we performed a series of experiments in each of which 60 randomly selected subjects are used with the two highest-intensities scans for each of the six basic facial expressions (i.e., each experiment includes 720 scans). The random selection of the subjects approximately guarantees that, in each experiment, the person- and gender-independency are preserved, and a good distribution of the subjects across the various ethnic groups. In each experiment, six *one*-vs.-*all* SVM classifiers, one for each expression, are trained and tested using the feature vectors $v_{\hat{f}}^{\text{expr}}$ and *10-fold cross-validation*. According to this, the 60 subjects are split into ten subsets, each containing 6 subjects. Of the 10 subsets, one subset is retained to test the model, and the remaining 9 subsets are used as training data, that is, the training set contained 54 subjects (648 scans), and the test set contained 6 subjects (72 scans). The ratio between positive and negative examples in the train and test subsets is equal to the ratio existing in the original data set. Using 10-fold cross-validation, training is repeated 10 times, with each of the 10 subsets used exactly once as the test data. Finally, the results from the ten steps are averaged to produce a single estimation of the performance of the classifier for the experiment. In this way, all observations are used for both training and test, and each observation is used for test exactly once. However, as pointed out in [10], since average recognition accuracies can vary from experiment to experiment, in order to permit a fair generalization and obtain stable expression recognition accuracies, we run 100 independent experiments and averaged the results (1000 train and test sessions in total).

Table 1 Most relevant features for each expression according to mRMR. Pairs (k, r) are reported, where k is the number of the sample point according to the numbering of Fig. 8(b), and r is the corresponding relevance (in percentage). Values in each column are ordered by decreasing relevance scores

Rank	Anger	Disgust	Fear	Happiness	Sadness	Surprise
	k, r	k, r	k, r	k, r	k, r	k, r
1	69, 100.0	57, 100.0	43, 100.0	52, 100.0	63, 100.0	7, 100.0
2	70, 92.3	58, 92.3	43, 81.2	42, 99.3	63, 96.9	6, 96.0
3	69, 77.7	68, 90.6	44, 77.1	19, 94.0	64, 96.8	8, 94.8
4	67, 76.7	7, 84.8	28, 75.6	51, 93.7	71, 96.0	5, 94.6
5	69, 75.0	5, 84.4	71, 74.0	52, 92.7	71, 93.0	8, 90.3
6	69, 73.8	49, 84.2	44, 73.7	42, 92.0	64, 88.7	4, 88.1

6.1 Feature selection

According to the feature selection analysis in Sect. 5, just the most relevant SIFT features identified with mRMR are used to perform 3D facial expression recognition. Table 1 summarizes, for the six basic expressions, the outcomes of mRMR by using the pair (k, r), where k is the index of the sample point according to the numbering reported in Fig. 8(b), and r represents the relevance (given in percentage) of the feature selected for the k sample point. The relevance value is obtained as the mutual information value returned by the mRMR algorithm, normalized by the mutual information of the most important feature. Actually, it should be noted that mRMR can select and use for classification either none, one or more than one of the 128 features of the SIFT descriptor at a particular sample point. From the table, it can be observed that sample points from which the most relevant SIFT features are extracted vary across expressions. In addition, the values reported in each column show as the relevance decreases with different trends from expression to expression. To account for this, only the features with a relevance not lower than 50% are retained among the 10112-dimensional SIFT feature vector, and these features are used to perform expression recognition. According to this, the first 18, 12, 8, 14, 16 and 20 features are used, respectively, for the *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* expressions. Considering all the expressions, SIFT features of just 33 sample points are used for recognition (see the Table in Fig. 9(b)). In addition, the results indicate that, with our approach, the large part of the information about facial expressions is conveyed by the local descriptors of sample points located in the mouth and cheek regions of the face, as shown in Fig. 9.



(a)

line	*#selected points*
1	5
2	6
3	4
4	4
5	2
6	9
7	2
8	1

(b)

Fig. 9 (**a**) The sample points for which at least one feature of the SIFT descriptor is selected for expression classification by the mRMR algorithm; (**b**) the number of selected points per line

6.2 Expression recognition

Using the selected mRMR features and SVM classification, 3D facial expression categorization is performed. The average recognition accuracies are computed by performing 100 independent runs, each including 10-fold cross-validation on the two highest intensities scans of 60 randomly selected subjects (720 scans per independent experiment).[1] The results are reported in Table 2 using the average *confusion matrix* as performance measure. Rows of the table are the

Table 2 Average confusion matrix (percentage values)

	An	*Di*	*Fe*	*Ha*	*Sa*	*Su*
An	**78.43**	2.15	3.50	1.24	13.01	1.67
Di	3.33	**77.05**	4.32	6.89	3.41	5.00
Fe	1.24	13.56	**67.50**	11.12	2.17	4.41
Ha	0.83	0.91	19.02	**77.42**	0.91	0.91
Sa	17.73	0.0	0.91	2.50	**78.86**	0.0
Su	0.0	6.01	2.68	0.0	0.0	**91.31**

[1] The identifiers of the 60 randomly selected subjects in each of the 100 experiments are publicly available upon request from the corresponding author.

Table 3 Comparison of this work with respect to Gong et al. (Gong) [10], Wang et al. (Wang) [33], Soyel and Demirel (Soyel) [29], and Tang and Huang (Tang) [30]. The average (AVG) expression recognition rates, in percentage, computed on all the six expressions and all the independent experiments are reported

	This work	Gong	Wang	Soyel	Tang
AVG	**78.43**	76.22	61.79	67.52	74.51

true expressions to classify, whereas columns represent the results of the classification. The overall recognition rate is equal to 78.43%. It can be observed that *surprise* is recognized with very high accuracy, whereas *fear* results in the expression more difficult to recognize (mainly confused with *disgust* and *happiness*). High confusion between *happiness* and *fear*, and between *sadness* and *anger* are also observed. Looking at Fig. 1, the difficulty to discriminate between *sadness* and *anger* can be motivated by the fact that these two expressions are very similar and can be discriminated mainly observing the differences in the eyes and eyebrows regions. However, these regions are typically acquired with noise by 3D scanners, so that we do not consider sampling points in these parts of the face (the *en–en* and *ex–ex* points are identified, but no sampling lines are considered between these points).

6.3 Comparative evaluation

In Table 3 the results of our approach are compared against those reported in [10] on a same experimental setting (the setting details are reported in Sect. 6.2).

For the approaches in [29, 30, 33], results are replicated from those reported in [10].[2] Some differences between the approaches listed in the table should be noted: Soyel and Demirel [29] use distances between manually identified landmarks (11 in total); Tang and Huang [30] use both distances between manual landmarks (83 in total), and neutral scans to normalize distances; Wang et al. [33] use manual landmarks (64 in total) to segment face regions; Gong et al. [10] obtain their best results subtracting neutral scans from depth region masks of the eyes and mouth. In comparison, our approach does not use neutral scans, but just relies on 9 automatically detected keypoints and 79 sample points derived on lines connecting the keypoints. In particular, it can be observed that our approach outperforms other solutions, with larger differences with respect to works that do not use neutral scans.

[2]We point out that the results reported in the original works of [29, 30, 33] are different from those summarized in [10]. However, these results are obtained for diverse experimental settings and thus do not permit a fear comparison. Due to this, we refer to the results in [10] that have been obtained for the same experimental setting of this work.

Table 4 The RMSE for x, y, and z coordinates, and the *mean* and *standard deviation* of the ADE (all measured in *mm*) for the 9 automatically located keypoints with respect to their manually annotated positions (BU-3DFE, high and highest expressions scans)

Landmark	RMSE			ADE	
	x	y	z	mean	stdev
prn	0.693	1.139	1.382	1.780	0.720
al_{left}	0.756	1.818	1.544	2.330	0.911
al_{right}	0.804	1.732	1.476	2.251	0.873
en_{left}	1.316	1.481	1.143	2.152	0.774
en_{right}	1.552	1.620	1.171	2.377	0.868
ex_{left}	3.225	2.070	2.126	4.132	1.455
ex_{right}	3.463	2.207	2.130	4.358	1.552
ch_{left}	3.546	2.860	2.639	4.970	1.732
ch_{right}	3.699	2.997	2.707	5.193	1.743

6.4 Keypoints positional accuracy

The set of 83 manually annotated landmarks provided with the BU-3DFE database (see Fig. 2) also includes the manual annotation of 8 of the 9 keypoints that are derived automatically with the approaches presented in Sect. 3. The nose tip is missing in the BU-3DFE, so this landmark was manually annotated in order to perform the experiment.[3] These manually annotated landmarks provide a ground-truth to evaluate the positional *Root Mean Square Error* (RMSE) of the 9 automatically located keypoints. Given the category k of a landmark ($k = 1, \ldots, 9$), the ground-truth coordinates of this landmark in the ith scan $l_i^k = (x_i^k, y_i^k, z_i^k)$, and the automatically derived estimation of the landmark $\hat{l}_i^k = (\hat{x}_i^k, \hat{y}_i^k, \hat{z}_i^k)$, the RMSE for the x-coordinate is defined as (similar definitions hold for the y- and z-coordinate):

$$\text{RMSE}(\hat{x}^k) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{x}_i^k - x_i^k\right)^2}, \tag{5}$$

where N is the number of scans used in the evaluation (i.e., N corresponds to the 1200 scans of the BU-3DFE database with the high and highest expression intensities used in the expression classification experiments). The *Absolute Distance Error* (ADE) in 3D is also computed as the Euclidean distance between a manually annotated landmark and its automatic estimation. The *mean* and *standard deviation* values of the ADE are reported in Table 4, together with the RMSE computed independently for the three coordinates x, y and z of each landmark.

[3]The annotation files for the nose tip are publicly available upon request from the corresponding author.

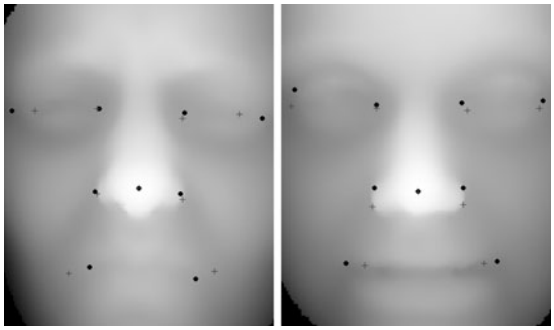


Fig. 10 Example range images showing the automatically detected keypoints (+) and their manually located positions (•)

It can be observed that the 9 keypoints were detected with different accuracies (two examples are shown in Fig. 10). The mean ADE for each of the 9 keypoints was less than 5.193 mm. The average ADE across all the 9 fiducial points was 3.283 mm.

The nose tip point (*prn*) was located most reliably in 3D (within 1.780 mm of its manual locations), followed by the inner corners of the eyes (*en–en*) points. The alare points were also located reliably for all the faces in the database. The corners of the mouth (*ch–ch*) were detected least reliably (the mean ADE for the left and right *cheilion* are equal to, respectively, 4.970 and 5.193 mm), preceded only by the outer corners of the eyes (*ex–ex*) with a mean ADE of 4.132 and 4.358 mm, respectively, for the left and right *exocanthion*.

The quite large errors observed for the *ex–ex* and *ch–ch* points with respect to the others can be motivated by the fact that manual annotations are carried out on 3D textured models so that the texture image largely influences the visual perception of the location of facial landmarks (especially for *ex–ex* and *ch–ch* due to the eyelashes and the lips, respectively). Differently, our completely automatic approach only relies on 3D data and would require, for a more fear evaluation of the positional accuracy of the keypoints, a ground-truth directly derived from the 3D face scans. In addition, the scans with the high and highest level of expression intensities were used for the detection of keypoints. Since these scans show very large expression variations, the regions of the face close to the mouth and eyes are modified to a large extent, thus making the location of the corresponding keypoints less accurate.

6.5 Facial expression retrieval

The proposed framework for 3D facial expression representation and classification can be also used to perform 3D *facial expression retrieval*. In a retrieval scenario, the idea is that the 3D scans of subjects with given facial expressions are used as target queries and the 3D scans of subjects that show the same facial expression as the target one are retrieved. For example, given a query scan of a subject with *happy* expression, a retrieval task permits to find all scans of subjects in the database with *happy* expression. This can have practical applications in different contexts, such as facial character animation, psychology studies, medical aesthetics, etc.

Table 5 Precision and recall values on the BU-3DFE (percentage values)

	An	*Di*	*Fe*	*Ha*	*Sa*	*Su*
Precision	77.23	77.30	68.93	78.07	80.17	88.39
Recall	78.43	77.05	67.50	77.42	78.86	91.31

In order to be applied for retrieval purposes, our framework for 3D facial expressions representation and classification requires that: First, the expression of the query scan is determined using SVM; Then, all the database scans that are classified in the same category of expression as the query are retrieved. Following this experimental setup, the retrieval results can be directly derived from Table 2. The standard *precision* and *recall* figures of merit that are used as performance measures of retrieval can be computed as:

$$precision = \frac{tp}{tp + fp}, \qquad recall = \frac{tp}{tp + fn}, \tag{6}$$

where tp, fp and fn are, respectively, the *true positives*, *false positives*, and *false negatives*. These values can be derived from Table 2, considering that the rows of the table are the true expressions, and the columns represent the corresponding classification. As a consequence, for each expression: (i) the true positives (tp) are on the principal diagonal of the confusion matrix; (ii) the false positives (fp) are the sum of non-diagonal values of the columns of the matrix; (iii) the false negatives (fn) are the sum of non-diagonal values of the rows of the matrix. According to this, precision and recall values are derived using (6) and are reported in Table 5.

To the best of our knowledge, derivation of the precision and recall performance is unique in the literature of 3D facial expression recognition. In particular, the results show an average performance close to 80% both for precision and recall.

7 Discussion and conclusions

In this paper, we proposed a completely automatic approach for person-independent facial expression recognition from 3D facial scans. The approach grounds on three main original contributions: (i) a solution to automatically detect 9

fiducial points of the face (*keypoints*) located in morphologically salient regions of the face; (ii) a local based description of the face that computes SIFT features on a set of sample points of the face derived starting from the 9 keypoints; (iii) a feature selection solution for the identification of the salient SIFT features. Using a multi-class SVM classification on a large set of experiments, an average facial expression recognition rate of 78.43% has been obtained for the six basic facial expressions on the publicly available BU-3DFE database. Remarkably, the comparative analysis shows that our completely automatic approach performs similarly or better than methods that use manual annotation of the face. The experiments also evidence that the proposed solutions for automatic identification of facial keypoints can locate fiducial points of the face with an accuracy compatible with the needs of an expression recognition approach. Finally, we also used the proposed face representation and classification framework to derive results of 3D facial expression retrieval. To this end, the 3D scans of subjects with a target facial expression are used as queries, with the goal to find all the 3D scans of subjects that show the same facial expression as the target scans. Recall and precision results show an average performance close to 80%.

The proposed approach can be extended in different directions in order to improve the facial expression recognition. On the one hand, methods to automatically identify further keypoints of the face could be defined so as to capture additional information in different regions of the face. In particular, in the current solution no keypoints are detected in the eyebrows region that, instead, is significantly modified by some expression changes (see for example the *disgust*, *fear* and *surprise* expressions in Fig. 1). On the other hand, the use of SIFT descriptor computed in correspondence with a set of sample points has proved its validity, but different solutions could be tried to capture the local 3D shape of the face. The idea is that different local face descriptors could be encompassed in our framework, provided that these local descriptors be sensible to variations originated by facial expressions. Finally, the retrieval perspective can be made more specific by developing a tailored retrieval approach. Following the common practice of retrieval applications, this could renounce to the classification framework in order to define a solution which uses feature extraction and matching between query and database scans.

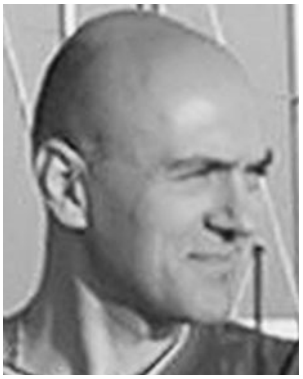
As future work, we plan also to evaluate the robustness of the proposed approach when applied to the scans of the BU-3DFE database with the lower and medium levels of expression intensities. These scans pose further challenges due to the small face variations induced by the low level of expression changes, and results on these subsets have not yet appeared in the literature of 3D facial expression recognition. Both the extensions to the proposed approach mentioned above could be tried to make our solution capable of accurately discriminating between small expression changes.

Acknowledgements The authors thank Professor Lijun Yin at the Binghamton University for making available the BU-3DFE database, and Alessandro Mannini at the University of Firenze for writing part of the code for automatic detection of facial keypoints. A preliminary version of this work appeared in [1]. The authors would like to thank the region Nord-Pas de Calais, France, for a visiting Professorship to Stefano Berretti under the program Ambient Intelligence. This research was also supported partially by the project FAR3D ANR-07-SESU-004.

References

1. Berretti, S., Ben Amor, B., Daoudi, M., del Bimbo, A.: Person-independent 3D facial expression recognition by a selected ensemble of SIFT descriptors. In: Proceedings of the 3rd Eurographics/ACM SIGGRAPH Symposium on 3D Object Retrieval, Norrköping, Sweden, pp. 47–54 (2010)
2. Berretti, S., Del Bimbo, A., Pala, P.: 3D face recognition using isogeodesic stripes. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 2162–2177 (2010)
3. Chalechale, A.: Content-based retrieval from image databases using sketched queries. Ph.D. thesis (2005)
4. Colombo, A., Cusano, C., Schettini, R.: 3D face detection using curvature analysis. Pattern Recognit. **39**(3), 444–455 (2006)
5. Do Carmo, M.P.: Differential Geometry of Curves and Surfaces. Prentice-Hall, Englewood Cliffs (1976)
6. Ekman, P.: Universals and cultural differences in facial expressions of emotion. In: Proceedings of the Nebraska Symposium on Motivation, Lincoln, NE, vol. 19, pp. 207–283 (1972)
7. Ekman, P., Friesen, W.V.: Manual for the Facial Action Coding System. Consulting Psychologist Press, Palo Alto (1977)
8. Farkas, L.G.: Anthropometry of the Head and Face. Raven Press, New York (1994)
9. Farkas, L.G., Munro, I.R.: Anthropometric Facial Proportions in Medicine. Thomas Books, Springfield (1987)
10. Gong, B., Wang, Y., Liu, J., Tang, X.: Automatic facial expression recognition on a single 3D face by exploring shape deformation. In: Proceedings of the ACM International Conference on Multimedia, Beijing, China, pp. 569–572 (2009)
11. Gupta, S., Markey, M.K., Bovik, A.C.: Anthropometric 3D face recognition. Int. J. Comput. Vis. **90**(3), 331–349 (2010)
12. Kakadiaris, I.A., Passalis, G., Toderici, G., Murtuza, N., Lu, Y., Karampatziakis, N., Theoharis, T.: Three-dimensional face recognition in the presence of facial expressions: an annotated deformable approach. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 640–649 (2007)
13. Lowe, D.: Distinctive image features from scale-invariant key points. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
14. Maalej, A., Ben Amor, B., Daoudi, M., Srivastava, A., Berretti, S.: Local 3D shape analysis for facial expression recognition. In: Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp. 4129–4132 (2010)
15. Maalej, A., Ben Amor, B., Daoudi, M., Srivastava, A., Berretti, S.: Shape analysis of local facial patches for 3D facial expression recognition. Pattern Recognit. **44**(8), 1581–1589 (2011)
16. Mayo, M., Zhang, E.: 3D face recognition using multiview key point matching. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, Genoa, Italy, pp. 290–295 (2009)

17. Mian, A.S., Bennamoun, M., Owens, R.: An efficient multimodal 2D–3D hybrid approach to automatic face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **29**(11), 1927–1943 (2007)
18. Mian, A.S., Bennamoun, M., Owens, R.: Keypoint detection and local feature matching for textured 3D face recognition. Int. J. Comput. Vis. **79**(1), 1–12 (2008)
19. Mpiperis, I., Malassiotis, S., Petridis, V., Strintzis, M.G.: 3D facial expression recognition using swarm intelligence. In: Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing, pp. 2133–2136 (2008)
20. Mpiperis, I., Malassiotis, S., Strintzis, M.G.: Bilinear models for 3-D face and facial expression recognition. IEEE Trans. Inf. Forensics Secur. **3**(3), 498–511 (2008)
21. Ohbuchi, R., Furuya, T.: Scale-weighted dense bag of visual features for 3D model retrieval from a partial view 3D model. In: Proceedings of the Workshop on Search in 3D and Video, Kyoto, Japan (2009)
22. Pandzic, I., Forchheimer, R.: MPEG-4 Facial Animation: The Standard, Implementation and Applications. Wiley, New York (2005)
23. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency max-relevance and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
24. Queirolo, C.C., Silva, L., Bellon, O.R., Segundo, M.P.: 3D face recognition using simulated annealing and the surface interpenetration measure. IEEE Trans. Pattern Anal. Mach. Intell. **32**(2), 206–219 (2010)
25. Ramanathan, S., Kassim, A., Venkatesh, Y.V., Wah, W.S.: Human facial expression recognition using a 3D morphable model. In: Proceedings of the IEEE International Conference on Image Processing, Atlanta, GA, pp. 661–664 (2006)
26. Rodriguez, J.J., Aggarwal, J.K.: Matching aerial images to 3-D terrain maps. IEEE Trans. Pattern Anal. Mach. Intell. **12**(12), 1138–1149 (1990)
27. Samir, C., Srivastava, A., Daoudi, M., Klassen, E.: An intrinsic framework for analysis of facial surfaces. Int. J. Comput. Vis. **82**(1), 80–95 (2009)
28. 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)
29. Soyel, H., Demirel, H.: Facial expression recognition using 3D facial feature distances. In: Proceedings of the International Conference on Image Analysis and Recognition, pp. 831–838 (2007)
30. Tang, H., Huang, T.S.: 3D facial expression recognition based on automatically selected features. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Anchorage, AK, pp. 1–8 (2008)
31. Vedaldi, A., Fulkerson, B.: VLFeat: an open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/
32. Venkatesh, Y.V., Kassim, A.A., Murthy, O.V.R.: A novel approach to classification of facial expressions from 3D-mesh data sets using modified PCA. Pattern Recognit. Lett. **30**(12), 1128–1137 (2009)
33. Wang, J., Yin, L., Wei, X., Sun, Y.: 3D facial expression recognition based on primitive surface feature distribution. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1399–1406 (2006)
34. 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, pp. 211–216 (2006)
35. Zheng, W., Tang, H., Lin, Z., Huang, T.S.: A novel approach to expression recognition from non-frontal face images. In: Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, pp. 1901–1908 (2009)



Stefano Berretti received the Lurea degree in Electronics Engineering and the Ph.D. in Information and Telecommunications Engineering from the University of Florence, Italy, in 1997 and 2001, respectively. In 2000, he received the post-Laurea degree in "Multimedia Content Design" from the University of Florence. Since 2002 he is an Assistant Professor at the University of Florence, where he teaches "Operating Systems" and "Fundamentals of Computer Programming" at the School of Computer Engineering. Since 2001 he also teaches "Database Systems" at the post-Doctoral school in "Multimedia Content Design" of the University of Florence. His scientific interests are pattern recognition, content based image retrieval, 3D object partitioning and retrieval, 3D face recognition.

Boulbaba Ben Amor received the M.Sc. degree in 2003 and the Ph.D. degree in Computer Science in 2006, both from Ecole Centrale de Lyon, France. Currently, he is an Associate Professor in Institut Telecom/Telecom Lille 1. He is also a member of the Computer Science Laboratory in University Lille 1 (LIFL UMR CNRS 8022). His research interests are mainly focused on statistical three-dimensional face analysis and recognition and facial expression recognition using 3D.
He is co-author of several papers in refereed journals and proceedings of international conferences. He has been involved in French and International projects and has served as program committee member and reviewer for international journals and conferences.

Mohamed Daoudi is a Full Professor of Computer Science in the Institut TELECOM; TELECOM Lille 1, LIFL (UMR 8022). He received the Ph.D. degree in Computer Engineering from the University of Lille 1 (USTL), France, in 1993 and Habilitation à Diriger des Recherches from the University of Littoral, France, in 2000. He was the founder of the MIIRE research group of LIFL (UMR 8022). His research interests include pattern recognition, image processing, three-dimensional analysis and retrieval and more recently 3D face recognition.
He has published more than 80 papers in refereed journals and proceedings of international conferences. He is the co-author of the book " '3D processing: Compression, Indexing and Watermarking" (Wiley, 2008).
He has served as a Program Committee member for the International Conference on Pattern Recognition (ICPR) in 2004 and the International Conference on Multimedia and Expo (ICME) in 2004 and 2005. He was a co-organizer and co-chair of ACM Workshop on 3D Retrieval 2010 and Eurographics 3D Object retrieval 2009. He has organized a

special session on 3D Face Analysis and Recognition at ICME 2008. He was an Associate Editor of the Journal of Multimedia (2006–2009). He is a frequent reviewer for IEEE Transactions on Pattern Analysis and Machine Intelligence and for IJCV, JMIV. His research has been funded by ANR, RNRT and European Commission grants. He is a Senior Member of IEEE.

Alberto del Bimbo is Full Professor of Computer Engineering, President of the Foundation for Research and Innovation, Director of the Master in Multimedia, and Director of the Media Integration and Communication Center at the University of Florence. He was the Deputy Rector for Research and Innovation Transfer of the University of Florence from 2000 to 2006. His scientific interests are multimedia information retrieval, Pattern recognition, image and video analysis and natural human–computer interaction. He has published over 250 publications in some of the most distinguished scientific journals and international conferences, and is the author of the monography "Visual Information Retrieval." From 1996 to 2000, he was the President of the IAPR Italian Chapter, and from 1998 to 2000, Member at Large of the IEEE Publication Board. He was the general Chair of IAPR ICIAP'97, the International Conference on Image Analysis and Processing, IEEE ICMCS'99, the International Conference on Multimedia Computing and Systems and Program Co-Chair of ACM Multimedia 2008. He is the General Co-Chair of ACM Multimedia 2010 and of ECCV 2012, the European Conference on Computer Vision. He is IAPR Fellow and Associate Editor of Multimedia Tools and Applications, Pattern Analysis and Applications, Journal of Visual Languages and Computing, and International Journal of Image and Video Processing, and was Associate Editor of Pattern Recognition, IEEE Transactions on Multimedia and IEEE Transactions on Pattern Analysis and Machine Intelligence.