Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

An efficient 3D face recognition approach based on the fusion of novel local low-level features

Yinjie Lei*, Mohammed Bennamoun, Amar A. El-Sallam

School of Computer Science and Software Engineering, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

ARTICLE INFO

Article history:
Received 19 March 2012
Received in revised form
27 June 2012
Accepted 29 June 2012
Available online 7 July 2012

Keywords:
3D face recognition
3D geometric feature
SVM
Feature-level fusion
Score-level fusion

ABSTRACT

We present a novel 3D face recognition approach based on low-level geometric features that are collected from the eyes-forehead and the nose regions. These regions are relatively less influenced by the deformations that are caused by facial expressions. The extracted features revealed to be efficient and robust in the presence of facial expressions. A region-based histogram descriptor computed from these features is used to uniquely represent a 3D face. A Support Vector Machine (SVM) is then trained as a classifier based on the proposed histogram descriptors to recognize any test face. In order to combine the contributions of the two facial regions (eyes-forehead and nose), both feature-level and score-level fusion schemes have been tested and compared. The proposed approach has been tested on FRGC v2.0 and BU-3DFE datasets through a number of experiments and a high recognition performance was achieved. Based on the results of "neutral vs. non-neutral" experiment of FRGC v2.0 and "low-intensity vs. high-intensity" experiment of BU-3DFE, the feature-level fusion scheme achieved verification rates of 97.6% and 98.2% at 0.1% False Acceptance Rate (FAR) and identification rates of 95.6% and 97.7% on the two datasets respectively. The experimental results also have shown that the feature-level fusion scheme outperformed the score-level fusion one.

1. Introduction

In the past two decades, 2D face recognition has been one of the most important and attractive research areas in computer vision [1,2]. However, pose and illumination variations have been the dominant factors which have hindered many practical applications of 2D face recognition systems. In order to overcome these limitations and inherent drawbacks, many researchers turned to 3D or surface facial information which is now commonly believed to have the potential to achieve a greater recognition accuracy than just 2D [3,4].

3D face recognition algorithms can be classified under different categories, e.g. according to the modality used, e.g. multimodal (RGB-D) vs. just 3D depth. Similarly to 2D, 3D face recognition algorithms can also be categorized into global-based, region-based and their hybrid-based (hybrid between global and local). Global recognition algorithms extract features AKA representations from the entire face. An example was proposed in [5], which extracts Gabor features from facial range images. However, one of the main challenges of 3D face recognition is the effect of facial expressions on the robustness and recognition accuracy. Global features are considerably affected by facial expressions. This contributed to the emergence of local-feature based algorithms, which commonly extract local features from the rigid parts of the face that are the least affected by facial expressions. A typical example in this category was proposed in our previous work [6], which showed that the nose and eyes-forehead regions contain reliable discriminating information for 3D face recognition. Hybrid-based approaches fuse global and local features/representations in order to enhance the recognition performance. Al-Osaimi et al. [7] proposed an expression-robust 3D face recognition algorithm which integrates local and global geometrical cues in a single compact representation of a facial scan. However, it was revealed to be hard to determine the weights of the global and local contributions when combining the two representations. To the best of our knowledge there is no existing theoretical work to prove that the integration of the two representations can definitely achieve better recognition results. On that basis, we opted to just use local based representations in this research.

Furthermore, in order to match/recognize two 3D surfaces which are defined in different coordinate systems, two major approaches are proposed. The first one requires the registration of the query (probe) surface to the reference (gallery) surface. One typical example is based on the Iterative Closest Point (ICP) algorithm [8,9] or one of its variants [10]. The ICP error has been used as a similarity measure to select the best matching surface

* Corresponding author. Tel.: +61 8 64882715.
E-mail addresses: yinjie@csse.uwa.edu.au (Y. Lei), bennamou@csse.uwa.edu.au (M. Bennamoun), elsallam@csse.uwa.edu.au (A.A. El-Sallam).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.06.023

for recognition. However, the ICP itself, as opposed to other algorithms, does not require any feature extraction or projection onto a higher dimensional space. It uses the whole surface, which makes it computationally expensive. It also sometimes fails because it does not converge to a global minimum. Another example is to use Principal Component Analysis (PCA) based approaches [11], which has also been extensively used in the 2D domain [12,13]. However, PCA-based approaches require a prior accurate fine registration. A second type of approaches to match/recognize two 3D surfaces defined in different coordinate systems is based on object-centric surface representations [14,15] which are used in an attempt to represent a 3D surface by a set of coordinate-independent features to eliminate the effect of the different coordinate systems. Consequently, one is able to develop statistical models which are learned from the features collected from the 3D surfaces of a training set to recognize any novel input. In order to build these statistical models, machine learning techniques have received increasing attention because of their great discriminating performance. On that basis, we propose a novel 3D face recognition approach based on the fusion of region-based low-level geometric features which are propagated to a set of SVM classifiers to perform face recognition.

The paper is organized as follows. Section 2 describes the related works and motivations. In Section 3, we provide the details of our proposed approach. In Section 4, extensive experiments are presented along with discussions. Finally, conclusions are given in Section 5.

2. Related work and overview

2.1. Related work

In its early year, most 3D face recognition approaches were based on global face representation/matching. Hesher et al. [16] presented a PCA-based approach to compute eigenvectors from 3D facial range images. Then a nearest neighbor classifier was used for identification. Achermann and Bunke [17] explored a 3D version of Hausdorff distance for 3D face recognition. They first computed two global facial representations which were based on point sets and voxel arrays respectively. 3D Hausdorff distances were then used as similarity measures for matching two faces. Later work showed that region-based 3D face recognition approaches perform better in achieving robustness under facial expressions. Chang et al. [18] segmented the 3D face into multiple sub-regions. The sub-regions which are located around the nose were matched individually and their matching scores were combined to determine the final recognition results. Faltemier et al. [19] also presented a region based method. In their work, a committee of facial regions were extracted and matched independently, and the final matching score was fused from the corresponding regions. They stated that the highest level of 3D face recognition had been achieved from combining 28 small regions. Zhong et al. [20] divided a 3D face into upper and lower regions, and used only the upper region without the mouth. Gabor filters were applied to extract features, and the centers from the filter response vectors were then learned by K-means clustering. Finally, the recognition results were obtained using a nearest neighbor classifier which was based on a learned visual codebook representation. Mian et al. [6] proposed an approach which automatically segmented the 3D face into different regions. A spherical face representation (SFR), scale-invariant feature transform (SIFT) based matching and a modified version of ICP algorithm were combined to achieve an expression robust $2D+3D$ face recognition. Queirolo et al. [21] presented an approach which used a simulated annealing-based algorithm for facial range image registration and measured the similarity of two faces using a Surface Interpenetration Measure (SIM). Four SIM values which correspond to four respective facial regions were then combined to obtain the final recognition results.

Many of the existing 3D face recognition approaches used high-level 3D facial features/representations, which were arguably insensitive to expression variations. Mian et al. [22] proposed a method to extract features around the neighboring areas of detected keypoints. These keypoints were defined as the areas of high shape variations and were extracted of a high repeatability. By fusing the 3D keypoint features with 2D SIFT features, they obtained 96.1% identification rate and 98.6% verification rate respectively on the FRGC v2.0 dataset. Wang et al. [23] developed a fully automatic 3D face recognition system using three types of local high-level features extracted from a Signed Shape Difference Map (SSDM) which was computed between two aligned 3D faces. Then a boosting algorithm was applied to select the most discriminative features to build three kinds of strong classifiers. They achieved a high recognition of 97.9% for verification and above 98% for identification on the FRGC v2.0 dataset. Berretti et al. [24] proposed a 3D face recognition system which partitions a 3D face into a set of isogeodesic stripes. Then a descriptor named 3D Weighted Walkthroughs was used to represent such strips, and a graph-based matching algorithm was used to match a pair of faces. Compared with high-level features/representations, the low-level geometric features provide less computational cost and are more reliable to represent the shape distribution of a 3D surface. Chua et al. [25] proposed a point-signature based method to represent the rigid regions of a 3D facial surface. Li et al. [26] proposed a robust 3D face recognition approach using a Sparse Representation for local geometric features and a pooling and ranking scheme was applied to choose higher-ranked expression-insensitive features. Gupta et al. [27] proposed an approach which automatically detected 10 anthropometric fiducial points using 2D and 3D face data. A stochastic pairwise method was used to calculate the 3D Euclidean and geodesic distances between all of the 10 points to perform 3D face recognition.

A large part of 3D biometric approaches are devised based on the registration between 3D surfaces using the Iterative Closest Point (ICP) algorithm or one of its modified versions. Some of the registration-based approaches achieve a satisfactory accuracy with 3D face, facial expression and ear recognition. However, these recognition approaches rely on brute force matching which affects their matching speed especially when the gallery size is large. In order to avoid brute force matching, an alternative is to use machine learning algorithms which train a set of statistical models, and recognize a novel input based on such models. Compared with brute force matching, machine learning based approaches provide efficient solutions to deal with a gallery of large scale. Hu et al. [28] used a geometric point distance based method for non-frontal facial expression recognition. They tested five different classifiers on the BU-3DFE dataset, and the highest recognition result was obtained using SVM. Sun et al. [29] investigated an efficient 3D dynamic facial expression recognition approach by establishing vertex correspondences across frames. They also proposed a spatiotemporal Hidden Markov Model (HMM) to learn the spatial and temporal information of a face. Based on their experimental results on a 3D dynamic face dataset, BU-4DFE, their proposed approach outperformed the approaches which use static 3D facial models. Maalej et al. [30] proposed a 3D facial expression recognition approach based on the shape analysis of the local facial patches. Their shape analysis algorithm computed the length of the geodesic path between local facial patches using a Riemanian framework. Besides, both SVM and multiboosting were used as classifiers, and they achieved 97.75% and 98.81% recognition rates respectively on the BU-3DFE dataset. Although machine learning techniques have been extensively studied in 3D facial expression recognition, their applications to the case of 3D face recognition is still in its infancy. This paper attempts to avoid brute force matching by using an SVM-based approach to perform 3D face recognition.

2.2. Overview of the proposed approach

In this paper, we first divide a 3D face into three different regions according to their distortions under facial expressions. We opted for the following face subdivision: rigid (nose), semi-rigid (eyes-forehead) and non-rigid (mouth) regions as shown in Fig. 2. In order to eliminate the effect of facial expression, we only take into account the rigid and semi-rigid regions. From each of these regions, we extract local surface descriptors. These descriptors are fused at the feature-level and at the score-level (Sections 3.5 and 3.6 for more details), and their recognition results are compared in Section 4.6.

We adopt low-level geometric features (Section 3.3.1) rather than high-level ones for the following reasons. First, the extraction of low-level features is computationally inexpensive involving only some basic computations e.g. distances and angles. On the contrary, high-level features usually require complicated mathematical transformations. Furthermore, every point on a 3D surface contributes equally to low-level features. This property enables that the magnitude of the change of these features to vary according to the magnitude of the change of the surface geometric distribution. Lastly, low-level features are invariant to scale, rigid motion and other similarity transformations as explained in Section 3.3. This property makes low-level features invariant to pose variations and independent of the coordinate systems.

We devise four types of region-based low-level geometric features (see Figs. 3 and 4) and apply them to 3D face recognition. Features that are collected from the same region are quantized into four types of histograms with a fixed dimension. Then those four histograms are concatenated to form a region-based descriptor (Section 3.3.2). In the rigid and semi-rigid facial regions, the extracted descriptors of the face of an individual at different instances are robust to facial expressions. Therefore by mapping such descriptors of a face, which is subject to various facial expressions, into a higher-dimensional feature space, the descriptors which belong to the same individual will cluster. We introduce SVM to find a hyper-plane in the higher dimensional space, which is then used to separate the cluster belonging to one individual from the others.

In order to combine the contributions of the rigid and semi-rigid facial regions, we apply both feature-level and score-level fusion. For feature-level fusion, we simply concatenate the two region-based descriptors/feature vectors and propagate them to SVMs for training and classification. For the score-level fusion, two region-based SVMs are trained separately, and their outputs are combined to obtain the final recognition results. In our approach, a likelihood normalization method is applied to optimally determine the combination weights.

3. Proposed approach

3.1. Facial scan preprocessing

The two largest available public human face datasets are used in this paper. They are the FRGC v2.0 (Face Recognition Grand

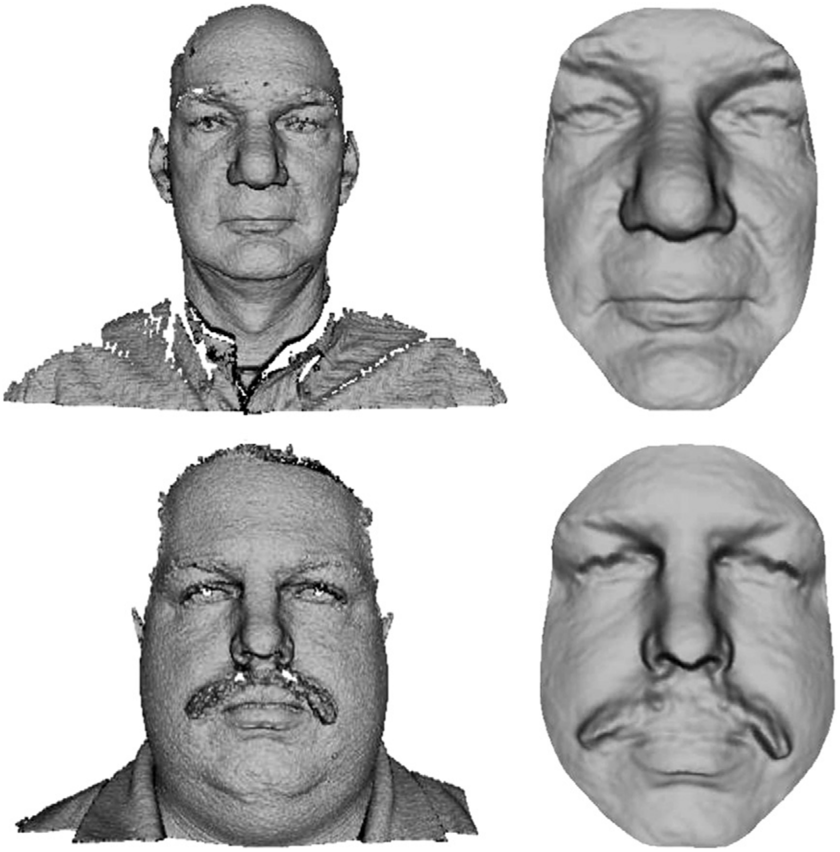

Fig. 1. An illustration of facial scan preprocessing, cropping and pose correction. The left column shows two raw facial scans with some obvious holes, spikes and outlier points. The right column shows the two respective processed range images. The surfaces have been smoothed to remove the spikes, and interpolated to fill in the holes. The pose has also been corrected. In addition, the nosetip is detected and shifted to the center of the image.

Challenge) dataset [31] and the BU-3DFE (Binghamton University 3D Facial Expression) dataset [32]. Section 4.1 provides a detailed description of these two datasets.

The 3D facial scans in FRGC v2.0 are represented by dense pointclouds and are stored in x, y and z matrices without any pre-processing. Some of the scans come with spikes and holes. In order to remove the spikes, we reduce/smooth the values of all three coordinates (x, y and z) of the outlier vertices according to the statistical information of the neighboring vertices and then smooth the whole mesh using a mean value filter. We then apply a bi-cubic interpolation along the three coordinate matrices to fill in the holes which often appear in the eye-brows or mouth areas. Since the original pointclouds are sampled at unordered (non-uniform) locations, the next pre-processing step is to build a uniform sampling pattern of the 3D facial scans in order to impose a fixed correspondence during the collection of the features (Section 3.3) across the different facial scans. This is accomplished by converting pointclouds to range images, which is a simple way of imposing a uniform sampling pattern on the 3D facial surfaces [26]. The range images are computed by interpolating at the integer x and y coordinates along the horizontal and vertical index respectively and determining the corresponding z coordinate as a pixel value.

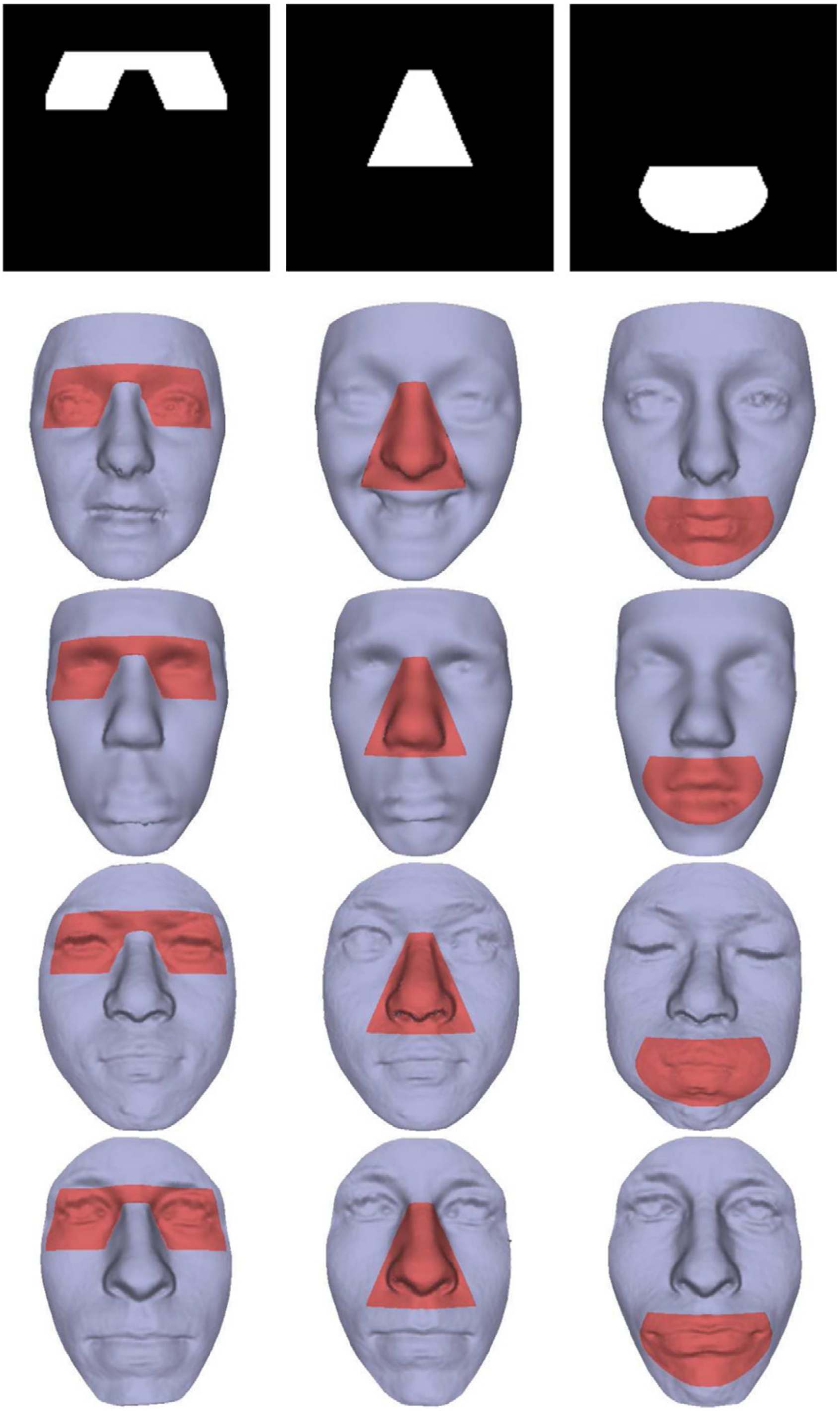

Fig. 2. Region based 3D facial representation. The 1st row shows the binary masks which are used to detect the semi-rigid, rigid and non-rigid regions of a face respectively. The 2nd and 3rd rows illustrate some extracted regions taken from the BU-3DFE dataset. The last two rows are extracted from FRGC v2.0 dataset.

The pixels in the range image are then re-sampled at a distance of 1 mm along both the x and y directions. A Gaussian filter is used after that to further smooth the range images.

In our previous work, we proposed an algorithm to crop facial scans and correct their pose to frontal views [6], which achieved sufficient accuracy. In our proposed approach, face cropping and pose correction are important for the facial regions which are extracted by means of pre-defined binary masks. The masks are developed by detecting the nosetip as a landmark. Thus, we first use a novel training free nosetip detection algorithm proposed in our previous work [33]. The range images are shifted to bring the nosetip to the center which is then set as the origin of the face coordinate frame, and follow the subsequent steps which are fully described in [6] and briefly described below. The detected nosetip is used to crop the 3D facial scans by eliminating the outlier points which are located at a distance of more than 80 mm from the nosetip. Pose correction is obtained by applying PCA on the cropped points to find their principal directions. This step is iterated until there is no further pose change taking place. Bi-cubic interpolation is then used to fill in the holes which occasionally appear during the pose correction caused by self-occlusion.

Note that the 3D facial models of BU-3DFE have already been preprocessed, cropped and pose corrected by the provider. Thus we only need to detect their nose tips and resample the facial scans into range images to guarantee the correspondence needed when extracting features from the different faces. Fig. 1 illustrates some of the pre-processing results. The left column shows two raw data from FRGC v2.0 dataset, and their corresponding processed range images are shown in the right column.

3.2. Region based 3D facial representation

As mentioned in Section 2.2, a facial scan is divided into different expression-sensitive regions. Both psychological findings and the 3D face recognition literature reported that the mouth area is the most affected under facial expression while the nose area is the least affected one. Based on that, we crop the nose (rigid region), the eye-forehead (semi-rigid region) and the mouth (non-rigid region) areas from the 3D facial scan for further processing. Many of the existing 3D face recognition approaches eliminate the non-rigid region in order to cope with facial expressions. Other approaches e.g. [21,6], provide a method to optimally decompose a 3D facial scan into different regions e.g. by detecting landmark feature points such as the inner and outside eye and mouth corners. Based on empirical results performed to assess the activity of each region, we devised three binary masks (shown in the top row of Fig. 2) to crop the facial regions. As shown in this figure (from left to right), the results of the decomposition of our binary masks are commensurate with the semi-rigid, the rigid and the non-rigid regions respectively.[1] Some examples of cropped facial regions are also shown in Fig. 2. In order to reduce the size of the extracted features, the cropped regions are sampled at uniform (x,y) intervals (2 mm in our case) and only the seed points are kept.

There are two advantages of using binary masks. First, complicated algorithms for landmark feature points detection are not required which results in a reduced computational cost. In addition, the cropped regions of different 3D facial scans will always contain the same number of vertices, which results in feature vectors of the same dimension (one of the requirements of SVM).

Our subsequent aim (Section 3.3) is to extract region-based low-level geometric features from 3D facial scans. For that purpose, we first converted the cropped range/depth information into pointclouds. The range image pixels are turned into (x,y,z) matrices, where x and y correspond to the vertical and horizontal indexes and the z value is taken from the corresponding depth value.

3.3. Extraction of 3D geometric feature functions

The low-level geometric features are directly computed from the spatial relationships of the 3D vertices without any complicated mathematical transformation [34]. The advantage of using spatial relationships can be explained as follows. Face pose variations or other rigid motions can only change the absolute spatial positions of the 3D vertices on a facial surface. The relative local spatial relationships among those vertices remains unaffected. Consequently, low-level geometric features which measure distances and angles between 3D vertices can be expected to remain invariant under pose variation or other similar rigid transformations. There are three qualities of a “good” feature representation. Namely, unambiguity, uniqueness and robustness [35]. A feature is unambiguous if different 3D faces yield different feature representations. A feature is unique if a 3D face is represented by a unique feature representation. Lastly, a feature is robust to facial expression if the magnitude of any variation caused by facial expression is much less than the magnitude of the change from one individual to another. In the subsection below, we propose four low-level geometric features which hold all of these qualities for the sake of an accurate and robust 3D face recognition.

3.3.1. Proposed 3D geometric feature functions

We represent each of the three facial regions (see Section 3.2) using multiple spatial triangles where one vertex is selected using the nosetip and the two other vertices are randomly picked from the corresponding local surface region. The example in Fig. 3 is shown on the rigid region. From these triangles, we develop four types of geometric features defined as follows:

1. A: corresponds to the angle between the two lines determined by the two random vertices and the nosetip.
2. C: is defined as the radius of the circumscribed circle to the triangle determined by the two random vertices and the nosetip.

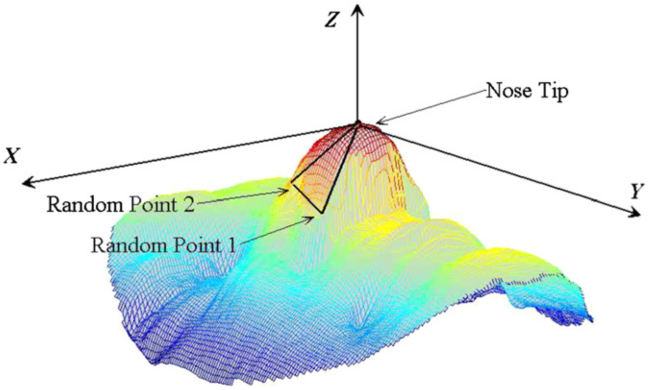

Fig. 3. An illustration of one of the triangles used for the extraction of our low-level geometric features. The triangle is defined by the nosetip and two random vertices picked from the selected facial region. The location of the nosetip is stored and used for all of the three facial regions.

[1] The code to generate the binary masks is available on the first author’s website at http://www.csse.uwa.edu.au/~yinjie/.

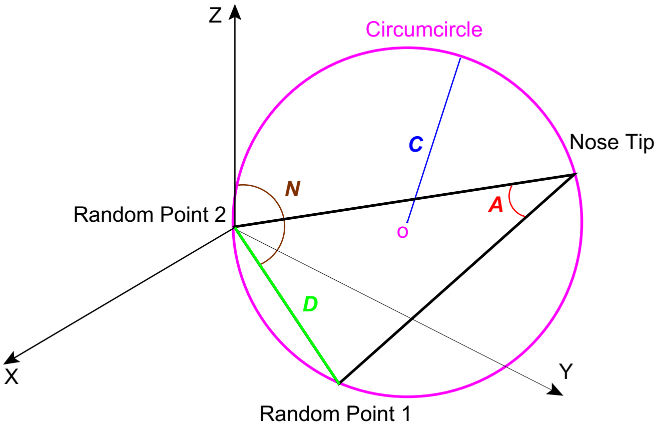

Fig. 4. An illustration describing the proposed four types of low-level geometric features.

3. *D*: is defined as the distance of the line between the two random vertices.
4. *N*: is defined as the angle between the line defined by the two random vertices and the *z*-axis.

An illustration of the four features is shown in Fig. 4. For the sake of clarity, we shifted the *z*-axis to one of the random vertices. The main strengths of our features can be summarized as follows: First, they are quick to compute and intuitively easy to understand. Second, these features vary dramatically between different individuals (unique) as proved in the following subsection. Third, our features are insensitive to facial expressions of the same individual (robust) as also proved in the following subsection. However, since two of the vertices which are used to define our triangles are randomly selected from the facial regions, the entries of the feature vector of two facial scans belonging to the same individual will not be in the same corresponding order. As a result, it is difficult to compare feature vectors of the same region. Hence we need to transform those unordered feature vectors into a comparable representation.

3.3.2. Proposed feature vector

Once the four types of features have been extracted, the following step is performed to overcome the random order problem mentioned in the previous subsection. In this work, we normalize the entries of each feature vector into $(-1,+1)$ and quantize them into histograms by counting how many entries fall into each of m bins.

We normalize and quantize each of the four vectors into histograms for the following reasons. First, we use a stochastic pairwise method to collect the four features. Therefore, there is no one-to-one correspondence of the entries of the feature vectors of two facial scans. Therefore, normalizing and quantifying feature vectors into histograms provides a better mean to describe the geometric distribution of the local surface. Furthermore, since the histograms are of the same dimension, the similarity between two surfaces can be computed by comparing their corresponding histograms. Second, because the size of the collected feature vectors is very large, we must compress them in order to reduce both the computational and the storage costs. For instance, with a cropped facial region of n vertices, $n/2$ vertices will still remain after using our seed-points resampling scheme (Section 3.2). Based on our spatial triangle representation, $(n^2-2n)/8$ triangles can be generated in total. Consequently, the dimension of each of the four feature vectors will also equal $(n^2-2n)/8$. Whereas the dimension of the histogram is only m (number of bins of the histogram), which is much less than $(n^2-2n)/8$. In addition, compared with those traditional feature compression algorithms, such as PCA, the histogram (with an optimal choice of m) is a "complete" representation which does not lose as much surface information.

Another important issue is to determine the histogram dimension. If the dimension is too large, the constructed feature histograms will be too sensitive to noise, expression or other variations. On the other hand, if the dimension is too small, all the elements will accumulate in some bins and thus yield to an insufficient discriminating information. Since it is difficult to theoretically determine the dimension, we empirically chose a set of dimensions and experimentally selected the optimal one (Section 4.2).

To demonstrate the robustness of our histogram descriptor, an illustration is provided in Fig. 5. We generate the four proposed histogram descriptors (all with a dimension of 180) from the rigid regions (nose) of two individuals, each under four facial expressions. This results in the generation of 16 histograms for each individual. The ones plotted in blue correspond to the first individual and the red ones correspond to the second individual. We can observe that for each individual, the histograms of the same feature are similar in terms of distribution. This is better illustrated in Fig. 6, where the histograms of the two individuals according to the different feature types are plotted together. For each feature type, it can be seen that the histograms belonging to the same individual are very similar, which results in a unique signature to discriminate one individual from another.

Finally, we concatenate all of the four histograms which are generated from the same region to form a region-based histogram descriptor. Consequently, a face can be represented by a rigid region histogram descriptor and a semi-rigid region histogram descriptor, each with a dimension of $m \times 4$.

3.4. Support vector machine classification

In this work, we adopt SVM for classification for its excellent discriminating performance. The basic idea of SVM is to map the feature vectors into a higher-dimensional space and then to find an optimal hyper-plane to separate one cluster from another by calculating the maximal margin. We first explain some relevant SVM concepts in this section.

3.4.1. The binary classification problem

SVM is a maximal margin classifier which performs classification by finding an optimal hyper-plane that maximizes the distance to the closest points in a higher-dimensional space. It was first designed to solve binary classification problems. Assume that we have q labeled training samples $x_k \in \mathbb{R}^D$, $k=1,\ldots,q$, which belong to two classes $y_k \in (+1,-1)$. SVM tries to find a hyper-plane by solving the following optimization problem

$$\min_{\omega,b,\xi} \left(\frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i \right)$$
$$\text{s.t.} \quad y_i(\omega^T\phi(x_i)+b) \geq 1-\xi_i, \quad \xi_i \geq 0, \tag{1}$$

where $C>0$ is a penalty parameter of the error term, ω is the coefficient vector, b is a constant and $\xi_i \geq 0$ is a parameter for handling non-separable data. In order to facilitate separation, each data is mapped by a function $\phi(x_i)$ into a higher-dimensional space. In practice, the function $\phi(x_i)$ is written in a form of a kernel function $K(x_i,x_j)=\phi(x_i)^T\phi(x_j)$, which calculates the dot product of two points in such a space. In our work, the non-linear Gaussian radial basis function (RBF) kernel is chosen because it

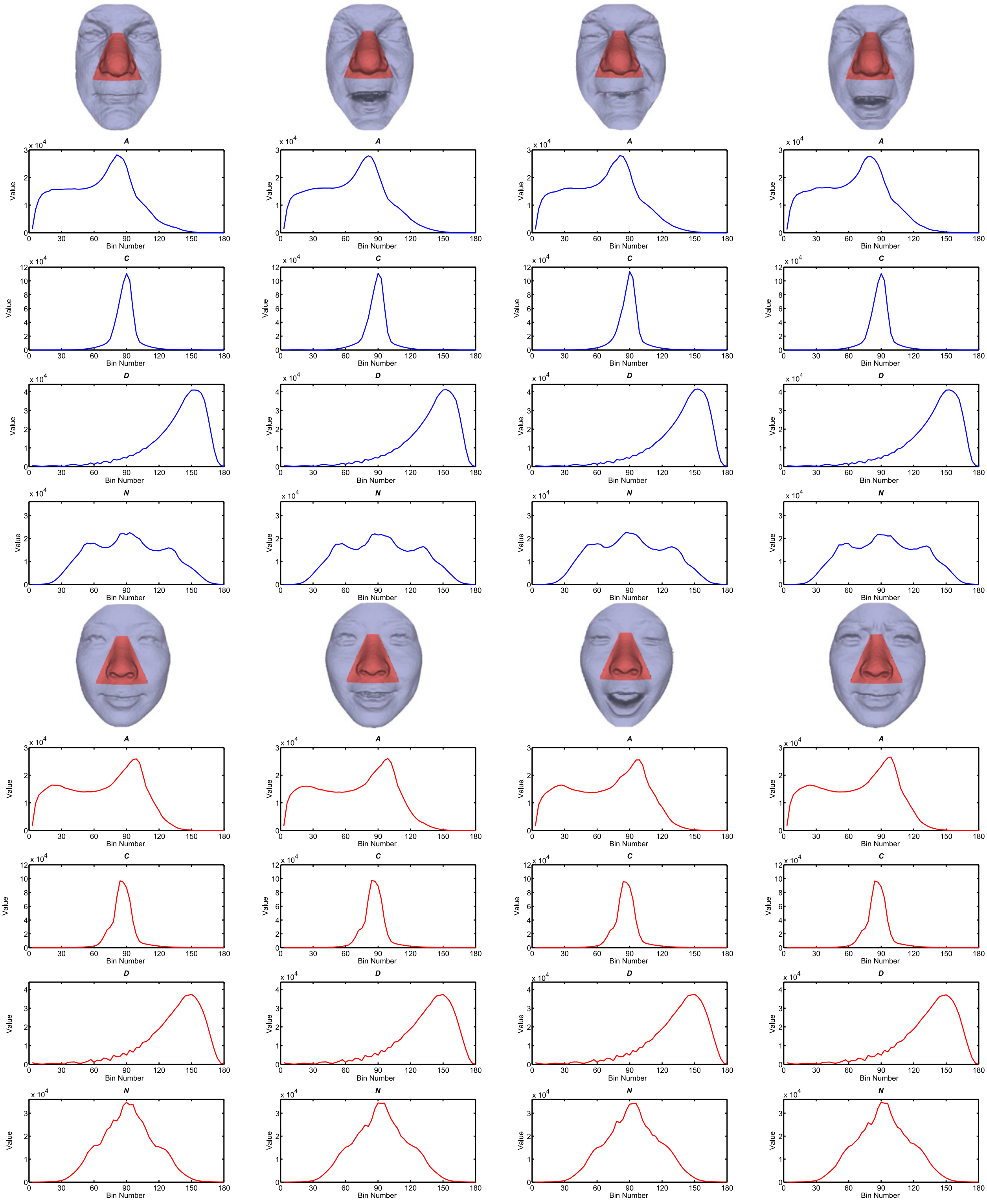

Fig. 5. The four proposed histogram descriptors collected from a rigid region of two individuals (each is under four different facial expressions). Notice that all of the histograms belonging to the same individual are very similar, whereas those of different individuals are dissimilar. A clearer illustration is given in Fig. 6.

has been shown to give better results compared to the linear and polynomial kernels [36]

$$K(x_i,x_j)=\exp(\gamma\|x_i-x_j\|), \quad \gamma>0, \tag{2}$$

where the kernel parameter γ along with the penalty parameter C need to be determined beforehand. The goal is to separate the data from two classes by a hyperplane that makes the distance to the support vectors maximized

$$f(x)=\omega\cdot x+b, \tag{3}$$

$$\omega=\sum_{\forall x_i\in S}\alpha_i y_i x_i, \tag{4}$$

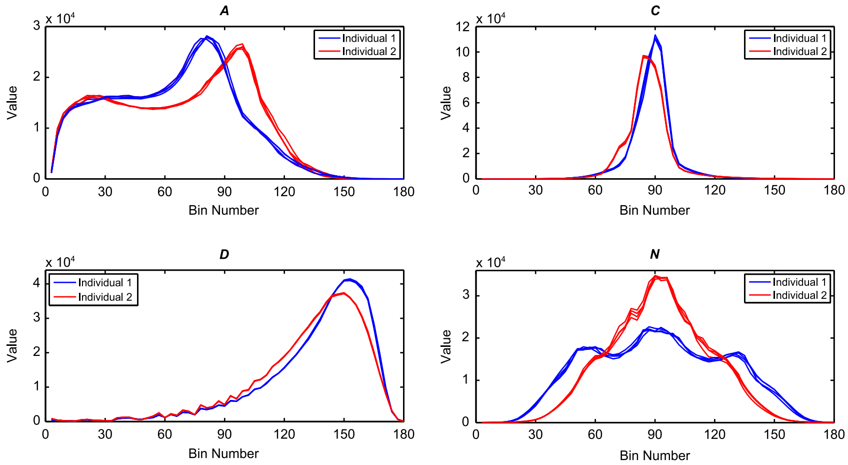

Fig. 6. An illustration to compare the histograms of the two individuals in Fig. 5 according to their types. It clearly demonstrates that the histograms which correspond to the same individual cluster together.

where S is a set of support vectors, and α_i is a trained weight of the corresponding support vectors. For each SVM, the hyperplane is calculated and stored to classify any novel input data using two classification schemes, one is a distance-based, and the other one is a probability-based.

1. Distance-based: by computing the sign of $d(x)$ which is a function of the right side of Eq. (3) as

$$y(x) = \begin{cases} +1 & \text{sign}(d(x)) = +1 \\ -1 & \text{sign}(d(x)) = -1 \end{cases}, \tag{5}$$

$$d(x) = \frac{\omega \cdot x + b}{\|\omega\|}, \tag{6}$$

   to perform binary classification, and the sign of $d(x)$ is the classification label of x, and $|d(x)|$ is the distance from x to the hyperplane. Intuitively, the farther distance from the data to the hyper-plane, i.e. a larger $|d|$, provides a more reliable classification result.
2. Probability-based: by mapping the real values of $f \in [-\infty, +\infty]$ onto probabilities, which can be obtained by training a sigmoid function which is defined as

$$P(y = +1|f(x)) = \frac{1}{1+\exp(Af+B)} \tag{7}$$

   where P is the mapped probability, y is the actual label of the input data, and A and B are two parameters estimated using the training set. The training set is fed back to the obtained SVMs, and a maximum likelihood estimation method [37] is applied on the output f_i and their actual labels y_i to estimate a pair of (A,B) which is used in the sigmoid function. For binary problems, a larger P indicates that the input data has a higher probability and will be classified as $+1$, and vice versa.

There is no theoretical analysis provided to compare the two methods, and it is hard to tell which one performs better than the other. However, the probability-based scheme provides a relatively more intuitive results since all its outputs are normalized between 0 and 1. In this paper the distance-based method has been applied to the feature-level fusion face recognition (Section 3.5) and the probability-based method has been used for the score-level fusion for the reasons explained below (Section 3.6).

3.4.2. *Multi-class classification problem*

Since SVM is originally designed for binary classification problems, there are two methods that are commonly used to solve multi-class problems: (1) The one-vs-all method, in which k SVMs are trained to classify k classes, and each SVM is responsible for the separation of the samples of one class (labeled $+1$) from all the other samples of the other classes of the training set (labeled -1). (2) The one-vs-one method, in which each SVM is responsible for classifying a pair of classes. The SVM is trained by treating the samples belonging to one class as positive (labeled $+1$) and those belonging to the other class as negative (labeled -1). As a result there are $k(k-1)$ trained SVMs, where k is the number of classes of the training set. Recent literature [36] shows that the two methods yield similar performance with respect to the classification quality. However, with respect to the training effort, the one-vs-all method only trains k SVMs compared with $k(k-1)$ that are trained by one-vs-one method. In this work, since we deal with a large classification problem, we take the one-vs-all method as it achieves a faster speed and a comparable performance to the one-vs-one method.

3.5. *Feature-level fusion*

We represent a 3D face by a semi-rigid region and rigid region histogram descriptors. How to fuse the recognition contributions from the two is an important issue. It is commonly believed that feature-level fusion can achieve a better performance than score-level fusion. However, the of feature fusion research is still in its infancy, and some existing work showed that score-level fusion provides better results [22]. In this work we present both feature-level and score-level fusion schemes. In this section, we first describe our feature-level fusion scheme.

3.5.1. *Training*

As feature-level fusion requires different features to be fused directly, in our case we concatenate the two region-based histogram descriptors of a 3D face into one feature vector to be fed as input to SVM. Then q SVMs are trained where q is the total number of individuals. Since the *RBF* kernel is selected in our case, the parameters C and γ need to be determined beforehand as mentioned in Section 3.4.1. We adopt a grid-search algorithm with 5-fold cross-validation to find the best pair of C and γ.

The pair which provides the minimum error during the cross-validation is selected.

3.5.2. Classification

We use a distance-based classification scheme because of its reduced training effort. Let x be a probe facial scan, the distance $d_k(x)$ is the kth element of the similarity vector $\overrightarrow{d}(x)$. Such distance value is computed according to Eq. (6) from the kth SVM, which is used to recognize the kth individual. Each vector is normalized on a scale of 0–1 using the min–max rule

$$\overrightarrow{d}'(x)=\frac{\overrightarrow{d}(x)-\min(\overrightarrow{d}(x))}{\max(\overrightarrow{d}(x)-\min(\overrightarrow{d}(x)))-\min(\overrightarrow{d}(x)-\min(\overrightarrow{d}(x)))}, \tag{8}$$

where $\overrightarrow{d}'(x)$ stands for the normalized probability vector. The operators $\min(\overrightarrow{d}(x))$ and $\max(\overrightarrow{d}(x))$ produce the minimum and maximum values of the vectors respectively.

For verification, given a classification threshold η

$$\textit{Accept} \text{ if } d'_k(x)+\eta>0$$

$$\textit{Reject} \text{ if } d'_k(x)+\eta\le 0. \tag{9}$$

If the threshold η is raised, the verification increases, but the false acceptance rate also increases, and vice versa.

For identification, the class label $y(x)$ of x is computed as follows:

$$y(x)=\arg\max_{1\le k\le q}(d'_k(x)), \tag{10}$$

where q is the number of individuals in the training set. The label is assigned with the individual which yields the largest $d'(x)$.

3.6. Score-level fusion

As opposed to feature-level fusion, score-level fusion requires the training of two sets of SVM, e.g. ***S***-SVMs and ***R***-SVMs which correspond to the semi-rigid and rigid regions respectively, and fuse the individual outputs to obtain the final classification results. Rather than concatenating the descriptors together, the region-based histogram descriptors are treated individually as feature vectors to train their corresponding SVMs. Then we adopt a weighted fusion method to optimally find their corresponding fusion weights of ***S***-SVMs and ***R***-SVMs. We use the probability-based scheme because the outputs generated from all SVMs are mapped and normalized between 0 and 1. The details of the score-level fusion are given in this section.

3.6.1. Training

Based on our region-based histogram descriptors, we train ***S***-SVMs and ***R***-SVMs which correspond to the semi-rigid and rigid regions respectively. The training process for both ***S***-SVMs and ***R***-SVMs is the same, therefore in the following we do not make any distinction between the two. The training also results in q SVMs to recognize q individuals. The outputs of each SVM are mapped and normalized into probabilities which take values between (0, 1) according to Eq. (7). As a result, a pair of sigmoid function parameters (A_k, B_k), $k=1,\ldots,q$, need to be estimated for each SVM as follows:

1. The training set is used to train q SVMs, and the grid-search under 5-fold cross-validation is used to find the optimal C and γ.
2. Feed all the labeled samples in the training set as inputs into the q SVMs, and use the kth SVM to generate a set of $(f_i,y_i)_k$, where i is the ith sample in the training set, and f_i is the decision value computed by Eq. (3) and y_i stands for the estimated label which takes a value of either -1 or $+1$.
3. A maximum likelihood estimation algorithm [38] is performed on $(f_i,y_i)_k$ to estimate the pair of (A_k,B_k).
4. Repeat step 2 and step 3 to find each of the q pairs of (A,B)s. As a result, q sigmoid functions are generated with their corresponding (A,B).

The above steps can be applied to train both ***S***-SVMs and ***R***-SVMs and their corresponding sigmoid functions. Consequently, for a probe facial scan x, two probability vectors $\overrightarrow{p}^s(x)=[p^s_1(x),\ldots,p^s_k(x),\ldots,p^s_q(x)]$ which corresponds to ***S***-SVMs and $\overrightarrow{p}^r(x)=[p^r_1(x),\ldots,p^r_k(x),\ldots,p^r_q(x)]$ which corresponds to ***R***-SVMs are computed and mapped by their corresponding sigmoid functions. Each of the vectors contains q elements with a sum of 1. At the location k, a larger value suggests a higher probability for the probe to be classified into the kth class. However, the outputs obtained by ***S***-SVMs are more reliable than those obtained by ***R***-SVMs, and this suggests to assign a different weight to each of them for fusion.

Let $\overrightarrow{w}^s$ and $\overrightarrow{w}^r$ be the weight vectors for the probability vectors $\overrightarrow{p}^s(x)$ and $\overrightarrow{p}^r(x)$ respectively. Each has q class-based elements: $\overrightarrow{w}^s=[w^s_1,\ldots,w^s_k,\ldots,w^s_q]$ and $\overrightarrow{w}^r=[w^r_1,\ldots,w^r_k,\ldots,w^r_q]$. We use a likelihood normalization method [37] to estimate $\overrightarrow{w}^s$ and $\overrightarrow{w}^r$.

Assume that there are n labeled samples $(x_1,\ldots,x_i,\ldots,x_n)$ in the training set, by feeding all the samples into ***S***-SVMs and ***R***-SVMs, two sets of probability vectors are obtained, e.g. $\overrightarrow{p}^s(x_i)$ and $\overrightarrow{p}^r(x_i)$, where $i=1,2,\ldots n$. $p^s_k(x_i)$ and $p^r_k(x_i)$ indicate the probabilities of the ith sample to be classified into the kth class obtained by ***S***-SVMs and ***R***-SVMs respectively. The class-based elements of $\overrightarrow{w}^s$ and $\overrightarrow{w}^r$ are computed by $w^s_k=\sum_{i=1}^n\sum_{k=1}^q p^s_k(x_i)/q\cdot\sum_{i=1}^n p^s_k(x_i)$ and $w^r_k=\sum_{i=1}^n\sum_{k=1}^q p^r_k(x_i)/q\cdot\sum_{i=1}^n p^r_k(x_i)$. In order to maintain the sum of the final probability vector to 1, the final weight vector $\overrightarrow{w}=w_1,\ldots,w_k,\ldots,w_q$ can be calculated by

$$w_k=w^s_k/(w^s_k+w^r_k). \tag{11}$$

As a result, the probabilities computed from ***S***-SVMs and ***R***-SVMs are fused to obtain the final probability vector

$$\overrightarrow{p}(x)=\overrightarrow{w}*\overrightarrow{p}^s(x)+(1-\overrightarrow{w})*\overrightarrow{p}^r(x), \tag{12}$$

where $*$ stands for the inner product operation, and each $p_k(x)$ in the probability vector $\overrightarrow{p}(x)=[p_1(x),\ldots,p_k(x),\ldots,p_q(x)]$ stands for the fused probability of the probe to be classified into the kth class.

3.6.2. Classification

Let x be a probe facial scan, the classification probability vector $\overrightarrow{p}(x)$ is computed according to Eq. (12).

For verification: given a classification threshold η

$$\textit{Accept} \text{ if } p'_k(x)>\eta$$

$$\textit{Reject} \text{ if } p'_k(x)\le\eta. \tag{13}$$

For identification: the class label $y(x)$ of x is computed as follows:

$$y(x)=\arg\max_{1\le k\le q}(p'_k(x)), \tag{14}$$

where q is the number of individuals in the training set. The label will be assigned with the class which yields the largest $p'_k(x)$.

4. Experimental results

We tested our proposed approach with a set of experiments on FRGC v2.0 and BU-3DFE datasets. In this section, we evaluate our proposed approach in both identification and verification modes, and present our recognition results.

4.1. Datasets description

FRGC is one of the largest available public domain 2D and 3D human face datasets. The FRGC v2.0 contains 4950 3D facial scans which are divided into training and validation partitions. The scans are captured by a Minolta Vivid series 3D scanner. The training partition contains 943 nearly frontal 3D facial scans (most of them are with a neutral expression) belonging to 273 individuals. This partition is used in Section 4.2 to select the optimal histogram dimension. The validation partition includes 4007 (Fall 2003 and Spring 2004) nearly frontal 3D facial scans of 466 individuals. There are 2410 facial scans out of 4007 which are in neutral expression, and the other 1597 are with non-neutral expressions. Two experimental subsets have been generated from the validation partition: (1) the 2410 neutral scans which belong to 466 individuals are used as a training set and 1000 non-neutral scans are used as a testing set. This is called "neutral vs. non-neutral"; (2) the 1597 non-neutral scans which belong to 374 individuals are used as a training set and 1000 neutral scans are used as a testing set. This is called "non-neutral vs. neutral".

BU-3DFE is another important public domain 3D face biometric dataset. It contains images and scans of 100 individuals (44 males and 56 females) with a large range of variety of age and ethnic/racial ancestries. Each individual has 25 ear-to-ear 3D facial scans, only one is in neutral expression and the rest 24 scans with six expressions, namely: Happiness, Anger, Fear, Disgust, Sadness and Surprise. Each expression has four levels of expression intensity. 1 and 2 levels are considered as low intensity while 3 and 4 levels as high intensity. Compared with FRGC v2.0, BU-3DFE provides a larger range of facial expressions at different intensities, and it is therefore more challenging when it comes to 3D face recognition. However, the total number of individuals of BU-3DFE is currently 100 only, which is very much less than the 466 individuals in FRGC v2.0. We also generated two subsets of experiment data from BU-3DFE: (1) for each individual, we select the low-intensity scans per expression in the training set and the high-intensity scans are included in the testing set. This is termed "low-intensity vs. high-intensity"; (2) we select the same number of training samples but with high-intensity per expression for each individual, and the low-intensity ones are included in the testing set. This is called "high-intensity vs. low-intensity". Consequently, both of the two experimental subsets contain 1200 samples for training and 1200 samples for testing.

4.2. Optimal histogram dimension

In this section, we aim to empirically find the optimal histogram dimension as described in Section 3.3.2 since this selection affects the recognition results. Two observations can easily be made. On the one hand, the larger the dimension of the histogram, the more sensitive is the recognition to noise, expression, and other variations. On the other hand, a small dimension will increase the chances of the histograms to concentrate on some specific bins which makes the descriptors more insensitive to variations between different individuals. Consequently, the recognition accuracy is affected.

The experiment is performed on the training partition of FRGC v2.0 as follows: we first randomly pick 150 individuals (out of 273), and since most of the facial scans in this partition are in neutral expression we randomly pick 300 scans from the selected individuals for testing and the rest for training. We conduct 20 experiments by repeating the random selection of training/testing data described above with histograms with dimensions ranging from 20 to 220 with an increment of 20. The experiment is run under the identification mode with the feature-level fusion to combine the rigid and semi-rigid regions and the average rank-1 Recognition Rates (RRs) and their 95% Confidence Intervals are reported in Fig. 7.

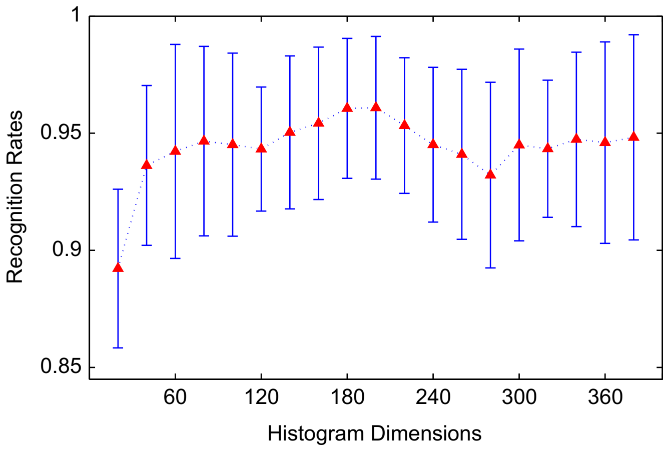

Fig. 7. The average rank-1 RRs and the 95% confidence intervals versus the histogram dimension using the feature-level fusion of rigid and semi-rigid regions.

As illustrated in Fig. 7, the average recognition rates increased when the dimensions are low (below 80). Then the recognition rate reached its peak in the range of 180–200. An obvious decrease of the rates occurred when the dimension of the histograms was above 220. Based on these observations, in the following experiments we use the region-based feature histogram with a dimension of 180.

4.3. Evaluation of the combinations of facial regions

In this work, we divide a facial scan into three regions, namely non-rigid (N), semi-rigid (S) and rigid (R) which correspond to mouth, eyes-forehead and nose areas respectively. In order to overcome the influence of facial expressions, we argue that the semi-rigid and rigid regions are relatively less sensitive to facial expressions. In order to validate our argument, a set of experiments are conducted on the validation partition of FRGC v2.0. First we randomly pick 200 individuals, and all their neutral facial scans are used for training. Meanwhile, another 500 scans with expressions which belong to the selected 200 individuals are used for testing. We conduct 20 experiments based on different combinations of these three facial regions by repeating the random selection of training/testing data as described above. Consequently, seven combinations of facial regions, namely N, S, R, $N+S$, $N+R$, $S+R$ and $N+S+R$, are generated and tested individually. The rank-1 RRs and their 95% confidence intervals are calculated and shown in Fig. 8. Two observations can be made: (1) The R region slightly outperforms the S region, whereas the N region is the most unreliable one; (2) The combinations which comprise the N region (N, $N+S$, $N+R$, and $N+S+R$) yield to lower rates compared to those without such a region (S, R, and $S+R$). The highest average recognition rate and the smallest confidence interval were obtained with the $S+R$ combination which exactly validates our argument that the semi-rigid and rigid regions are less sensitive to facial expressions.

4.4. Evaluation of the combinations of features

In order to investigate the reliability of our proposed low-level geometric features, a set of experiments were conducted based on 15 different combinations of all the four types of features. As discussed above, the histogram dimension is also set to 180, and $S+R$ regions are combined based on feature-level fusion. The rank-1 RRs and their 95% confidence intervals by repeating 20

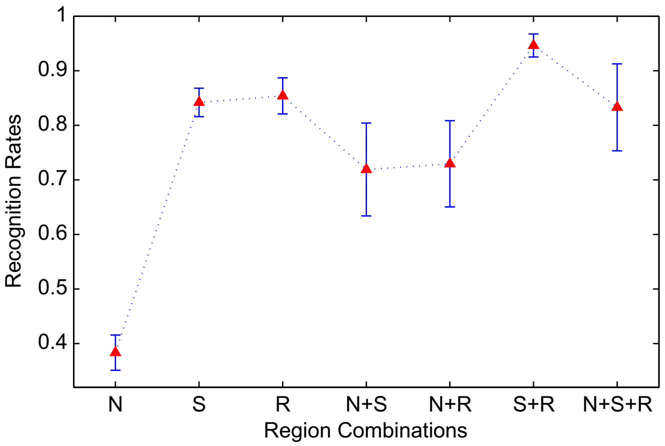

Fig. 8. The average rank-1 RRs and the 95% confidence intervals versus the combination of the three facial regions. The feature-level fusion is used to combine the different regions.

Table 1
The average rank-1 RRs and the 95% confidence intervals combining all of the four types of low-level geometric features, which are obtained by feature-level fusion of *S* and *R* regions by repeating 20 times the random selection of training/testing data.

Combinations	Rank-1 RRs (%)	Confidence intervals (%)
A	80.9	[77.8–84.1]
C	73.8	[69.3–78.3]
D	86.3	[83–89.6]
N	85.1	[82.2–87.9]
A+*C*	84.8	[80.6–89]
A+*D*	91.1	[87.4–94.9]
A+*N*	92.3	[90.2–94.5]
C+*D*	90.2	[87.8–92.7]
C+*N*	88.9	[86–91.9]
D+*N*	90.7	[88.7–92.7]
A+*C*+*D*	92.4	[90.3–94.5]
A+*C*+*N*	91.6	[89–94.3]
A+*D*+*N*	93.7	[91–96.3]
C+*D*+*N*	93.2	[90.7–95.8]
A*+*C*+*D*+*N	***94.3***	[***92.2–96.4***]

times the random selection of training/testing data described in Section 4.3 are shown in Table 1. These results show that, for a single feature, the feature ***D*** gives the highest performance. Feature ***C*** yields to an inferior performance compared with the other three. However, the overall best performance is achieved by combining all four types of features with an average recognition rate of 94.3%. Based on these experimental results, two observations can be made. First, the overall recognition rate using any of the feature types is satisfactory. Furthermore, the highest recognition performance is attributed to the fusion of all four types of features.

4.5. *Comparison of different classifiers*

In order to solve the non-linearly separable problem, the “kernel trick” is introduced to map the data from the original feature space into a high dimensional space, in which the mapped data can be expected to become more discriminative and separable by a hyper-plane. In order to demonstrate the superiority of the non-linear SVM, a comparison between a modified LDA algorithm [39,40] (the modified LDA algorithm resolves the singularity problem caused by small sample-sized training data [41,42]) and three SVM-based algorithms (Linear-SVM, Polynomial-SVM and *RBF*-SVM) is presented in Table 2. All experiments in this section are performed according to the same experimental setups of Section 4.3.

Table 2
Comparison of the different classifiers. The average rank-1 RRs and their confidence intervals are also obtained by feature-level fusion of *S* and *R* regions by repeating 20 times the random selection of training/testing data.

Classifiers	Rank-1 RRs (%)	Confidence intervals (%)
Modified-LDA	77.8	[74.1–81.6]
Polynomial-SVM	81.8	[78.2–85.4]
Linear-SVM	93.7	[91–96.4]
***RBF*-SVM**	***94.5***	[***91.9–97.2***]

The results provided in Table 2 show that all three SVMs achieve a better performance in terms of average recognition accuracy and stability (i.e. small confidence intervals) compared with the modified LDA algorithm mentioned above. Another observation is that the *RBF*-SVM slightly outperforms the Linear-SVM (i.e. a higher average recognition accuracy). This suggests that some of the original non-linearly separable data become more linearly separable following the mapping onto a high dimensional space using the *RBF* kernel, and thus the recognition accuracy is improved. However, based on our experimental results the Polynomial-SVM is not a suitable choice.

4.6. *Recognition results*

Cumulative Match Characteristic (CMC) curves from our proposed approach are provided in Fig. 9. Both feature-level and score-level fusion to combine the *S* and *R* regions are tested on FRGC v2.0 and BU-3DFE. The rank-1 identification rates provided by the feature-level fusion in the case of “neutral vs. non-neutral” and “non-neutral vs. neutral” of FRGC v2.0 are 95.6% and 96.7% respectively, which outperform the score-level fusion by 1.2% and 1.4% respectively. The results obtained on BU-3DFE are almost identical to FRGC v2.0, and the rank-1 identification rates for “low-intensity vs. high-intensity”, “high-intensity vs. low-intensity” are 97.7% and 98.7% (in the case of feature-level fusion) respectively versus 95.3% and 96.9% (in the case of score-level fusion). These results indicate that feature-level fusion outperforms score-level fusion which confirm previous works e.g. [43].

Fig. 10 illustrates the Receiver Operation Curves (ROCs) on the two datasets. Similar to the identification cases, both feature-level and score-level fusions combining the *S* and *R* regions are also tested. At 0.1% FAR, when our proposed approach is tested on FRGC v2.0 it achieves 97.6% and 98.1% in the case of feature-level fusion for “neutral vs. non-neutral” and “non-neutral vs. neutral” respectively. This outperforms score-level fusion by 1% and 0.9% respectively. On BU-3DFE, we obtain a feature-level fusion based verification rates of 98.2% and 99% in the case of “low-intensity vs. high-intensity” and “high-intensity vs. low-intensity”, which outperform score-level fusion by 0.8% and 0.7% respectively.

Although BU-3DFE is expression-richer than FRGC v2.0, the recognition results on BU-3DFE slightly outperform the FRGC v2.0 results. The differences in the recognition results occur because of the following two reasons. The first reason relates to some incorrect automatic nosetip detection and pose correction of the FRGC v2.0 facial scans. The results reported throughout this paper include these facial scans and no intervention was attempted to correct their pose manually. The performance of the approach can further be improved by performing an iterative nosetip detection while pose-correcting the 3D face. A more accurate nosetip detection will result in a more accurate pose-correction for 3D faces. The second reason is due to the different sizes of the the two datasets (number of individuals). The FRGC v2.0 contain 466 individuals in total, whereas the BU-3DFE comprises only 100

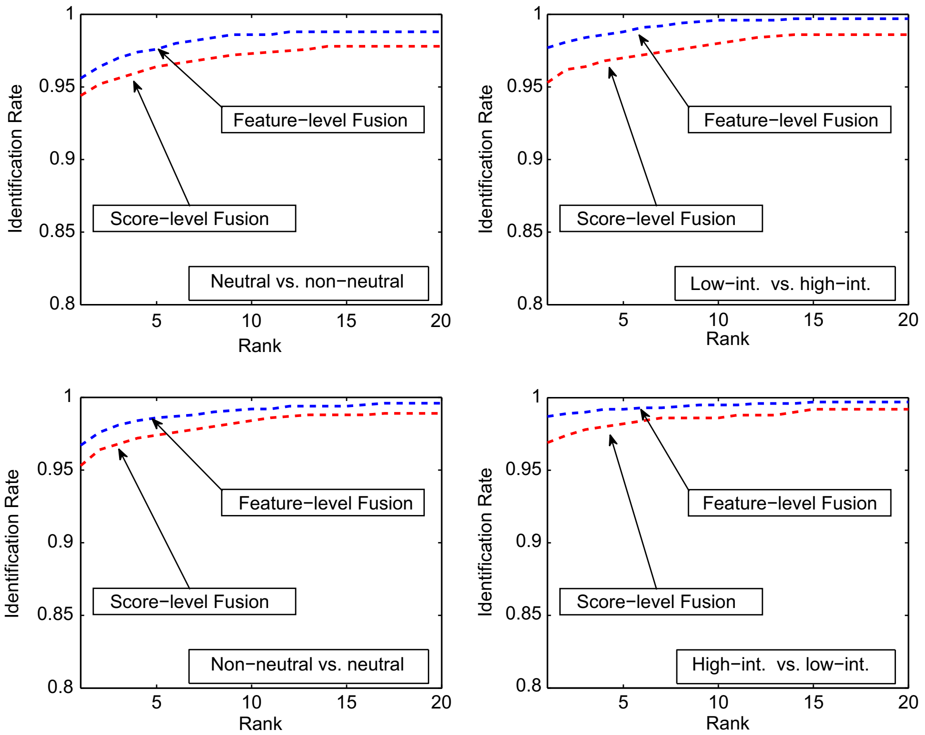

Fig. 9. Face identification results on FRGC v2.0 (first column) and BU-3DFE (second column). The rank-1 identification rates of combining $S+R$ regions by feature-level fusion on the two most challenging subsets, “neutral vs. non-neutral” of FRGC v2.0 and “low-intensity vs. high-intensity” of BU-3DFE, are 95.6% and 97.7% respectively.

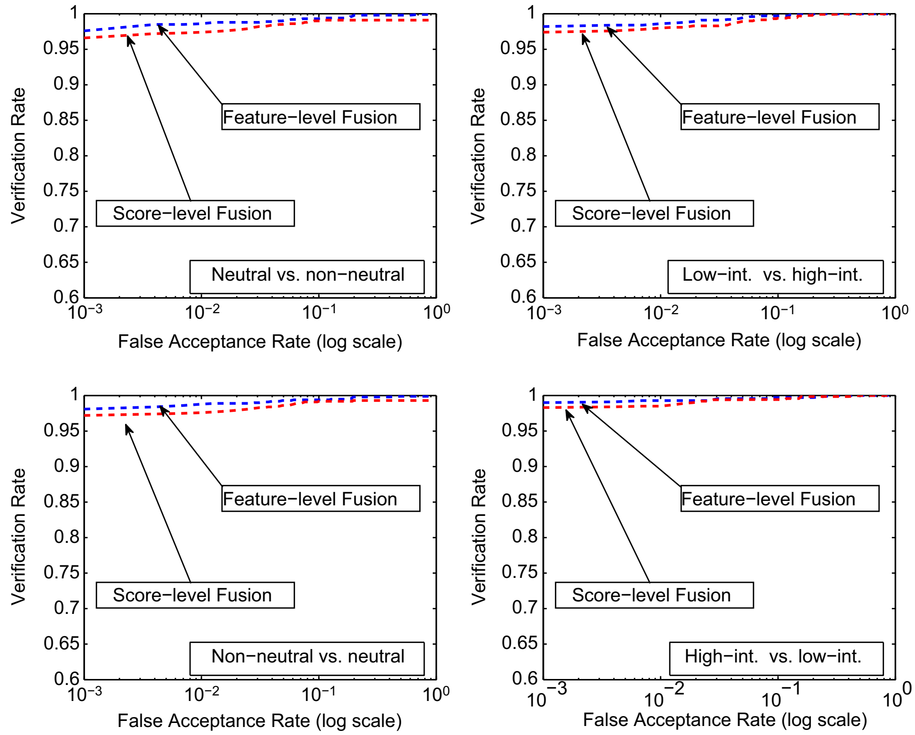

Fig. 10. ROC curves of the proposed approach on FRGC v2.0 (first column) and BU-3DFE (second column). The verification rates combining $S+R$ regions by feature-level fusion at 0.1% FAR on the two most challenging subsets, “neutral vs. non-neutral” of FRGC v2.0 and “low-intensity vs. high-intensity” of BU-3DFE, are 97.6% and 98.2% respectively.

individuals. It is commonly believed that a larger number of classification categories will make the recognition more challenging.

In this section a comparison of feature-level and score-level fusion has been performed using our proposed preprocessing steps and features/representations. However, it is worth emphasizing that the proposed feature-level fusion outperforms score-level fusion under our specific preprocessing steps (outlined in Section 3.1) and when adopting these particular proposed features. The adoption of different preprocessing steps, features and matching/recognizing algorithms may lead to a different outcome. Table 3 provides a comparison with

Table 3
Comparison of results for "neutral vs. non-neutral" face recognition using a gallery of 466 individuals from the FRGC v2.0 dataset.

Method	Modalities	No. of probes	Veri. rates (%)
Wang et al. [23]	3D	1597(466)	97.7
Mian et al. [6]	2D+3D	1597(466)	98.3
Berretti et al. [24]	3D	1538(466)	91.4
Mian et al. [22]	2D+3D	1597(466)	96.6
Al-Osaimi et al. [44]	2D+3D	500(466)	97.9
This paper	3D	1000(466)	***97.6***

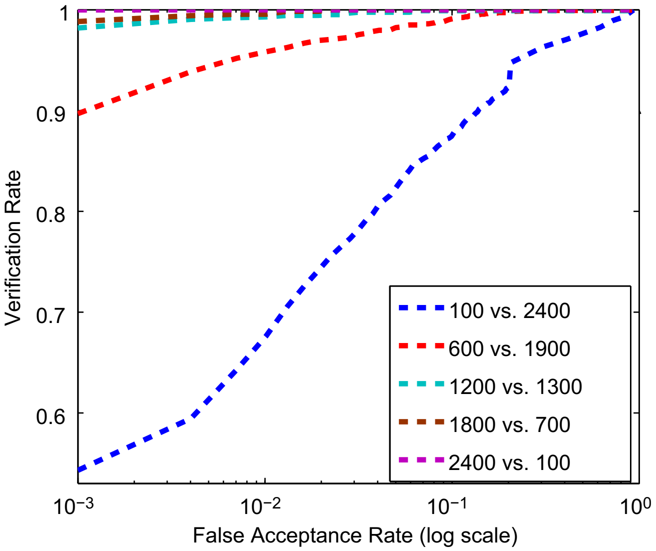

Fig. 11. ROC curves for different ratio of training/testing samples on BU-3DFE dataset.

existing state of the art techniques under different setups in the case of "neutral vs. non-neutral" of FRGC v2.0 dataset. The performance of our feature-level fusion approach outperforms one feature-level fusion approach proposed in our previous work [22], and another more recent approach [24]. The same table shows that the approach in this paper performs slightly less than our approach in [6], and its performance is very close to the approaches of [23] and our approach in [44]. The slight superior performance of these approaches may be attributed to better preprocessing steps [23], the adoption of complex features/representations [6,23,24] or the advantage of multi-modality (the combination of 2D and 3D data) [6,22,44]. Instead, our approach is only based on a simple geometric feature with a single modality (i.e. 3D only).

4.7. Sensitivity to the ratio of training/testing samples

In this section, we investigate the performance of the proposed approach in terms of the variation of the ratio of training/testing samples. For that purpose, we generate four randomly data subsets from BU-3DFE. We use BU-3DFE (as opposed to FRGC v2.0 dataset) because it provides the same number of facial scans per individual. In these four subsets, for the 100 individuals we vary the number of training samples between 100 (i.e. 1 scan per individual) and 2400 (i.e. 24 scans per individual). We vary the number of testing samples from 2400 to 100 for the 100 individuals. The resulting overall face verification rates using our proposed approach on the four subsets are shown in Fig. 11.

As expected, the proposed approach achieves the lowest verification rate (54.9% at 0.1% FAR) for the dataset which only contains 100 training samples and 2400 testing samples. The second subset which contains 600 training samples and 1900 testing samples provides a relatively higher verification rate of 89.9% at 0.1% FAR. The extremely high verification rate (100%) is obtained for the dataset which contains 2400 samples for training and only 100 testing samples. This experimental result clearly demonstrates that the insufficiency of training samples will challenge the recognition results of our SVM based 3D face recognition approach. However, our approach achieves reasonable rates for the first and also the most challenging dataset, which clearly demonstrates the robustness of our proposed approach when dealing with the 3D face recognition problem.

5. Conclusion

In this work we proposed a local geometric feature and SVM based 3D face recognition approach and tested its performance on FRGC v2.0 and BU-3DFE datasets. A facial scan is divided into three regions based on their sensitivity to facial expression. Four types of local geometric features are extracted from the rigid and semi-rigid regions. The local geometric features that are collected from the same region are then converted into a region-based histogram descriptor. Our tests have shown that when applying SVM and the fusion (at both feature- and score-level) of our region-based histogram descriptors (extracted from the rigid and semi-rigid regions), we achieve a good recognition performance on both FRGC v2.0 and BU-3DFE (both above 97.5% at 0.1% FAR verification rates for the most challenging experimental subsets). Based on experimental tests, one can make the following two observations. The first is that feature-level fusion has a better performance compared to score-level fusion in both of the identification and verification modes. The second one is that the fusion of the rigid and semi-rigid regions of a 3D facial scan contains the most reliable discriminating features which are important to achieve a robust 3D face recognition in the presence of facial expressions. To the best of our knowledge, this paper is the first of its kind to be based on the training of low-level geometric descriptors and the adoption of a machine learning approach. Our algorithms have been tested on the two largest public domain datasets, and our recognition (identification and verification) results, which are comparable or even outperform the state of the art one-to-one matching techniques. We believe that our algorithms can be subject to improvements. Our facial cropping scheme relies on fixed binary masks for all the facial scans. In our future work, we will propose a more precise algorithm to detect and crop facial regions.

Acknowledgment

The first author would like to thank the China Scholarship Council (CSC) for its financial support towards a PhD scholarship. We would also thank Australian Research Council (ARC) for its financial support under the Grant DP110102166.

References

[1] W. Zhao, R. Chellappa, P. Philips, A. Rosenfeld, Face recognition: a literature survey, ACM Computing Surveys (CSUR) Archive 35 (4) (2003) 399–458.
[2] A. Abate, M. Nappi, D. Riccio, G. Sabatino, 2D and 3D face recognition: A survey, Pattern Recognition Letters 28 (14) (2007) 1885–1906.
[3] K. Bowyer, K. Chang, P. Flynn, A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition, Computer Vision and Image Understanding 101 (1) (2006) 1–15.
[4] C. Xu, Y. Wang, T. Tan, L. Quan, Depth vs. intensity: which is more important for face recognition?, in: Proceedings of 17th International Conference on Pattern Recognition (ICPR 2004), vol. 4, 2004, pp. 342–345.
[5] Y. Wang, C. Chua, Y. Ho, Facial feature detection and face recognition from 2D and 3D images, Pattern Recognition Letters 23 (2002) 1191–1202.
[6] A. Mian, M. Bennamoun, R. Owens, An efficient multimodal 2D and 3D hybrid approach to automatic face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (11) (2007) 1927–1943.

[7] F. Al-Osaimi, M. Bennamoun, A. Mian, Integration of local and global geometrical cues for 3D face recognition, Pattern Recognition 41 (2008) 1030–1040.

[8] P. Besl, N. Mckay, Reconstruction of real-world objects via simultaneous registration and robust combination of multiple range images, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 239–256.

[9] Y. Chen, G. Medioni, Object modeling by registration of multiple range images, IEEE Transactions on Pattern Analysis and Machine Intelligence 3 (1991) 2724–2729.

[10] R. Sablatnig, M. Kampel, Model-based registration of front- and backviews of rotationally symmetric objects, Computer Vision and Image Understanding 87 (1–3) (2002) 90–103.

[11] T. Russ, C. Boehnen, T. Peters, 3D face recognition using 3D alignment for PCA, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[12] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.

[13] M. Kirby, L. Sirovich, Application of the KL procedure for characterization of human faces, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1) (1990) 103–108.

[14] A. Mian, M. Bennamoun, R. Owens, Three-dimensional model-based object recognition and segmentation in cluttered scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (2006) 1584–1601.

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

[16] C. Hesher, A. Srivastava, G. Erlebacher, A novel technique for face recognition using range imaging, in: Proceedings of Seventh International Symposium on Signal Processing and Its Applications, vol. 2, 2003, pp. 201–204.

[17] B. Achermann, H. Bunke, Classifying range images of human faces with Hausdorff distance, in: Proceedings of 15th International Conference on Pattern Recognition, Barcelona, Spain, vol. 2, 2000, pp. 809–813.

[18] K. Chang, W. Bowyer, P. Flynn, Multiple nose region matching for 3D face recognition under varying facial expression, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1695–1700.

[19] C. Faltemier, K. Bowyer, P. Flynn, A region ensemble for 3-D face recognition, IEEE Transactions on Information Forensics and Security 13 (1) (2008) 62–73.

[20] C. Zhong, Z. Sun, T. Tan, Robust 3D face recognition using learned visual codebook, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.

[21] C. Queirolo, L. Silva, O. Bellon, 3D face recognition using simulated annealing and the surface interpenetration measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2) (2010) 206–219.

[22] A. Mian, M. Bennamoun, R. Owens, Keypoint detection and local feature matching for textured 3D face recognition, International Journal on Computer Vision 79 (1) (2007) 1–12.

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

[24] S. Berretti, A. Del Bimbo, P. Pala, 3D face recognition using isogeodesic stripes, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (12) (2010) 2162–2177.

[25] C. Chua, F. Han, Y. Ho, 3D human face recognition using point signature, in: Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Frenoble, France, 2000, pp. 233–239.

[26] X. Li, T. Jia, H. Zhang, Expression-insensitive 3D face recognition using sparse representation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2575–2582.

[27] S. Gupta, M. Markey, A. Bovik, Anthropometric 3D face recognition, International Journal of Computer Vision 90 (3) (2010) 331–349.

[28] Y. Hu, Z. Zeng, L. Yin, X. Wei, J. Tu, T.S. Huang, A study of non-frontal-view facial expressions recognition, in: Proceedings of 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[29] Y. Sun, X. Chen, M. Rosato, L. Yin, An evaluation of multi-model 2D+3D bimetrics, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 40 (3) (2010) 461–474.

[30] A. Maalej, B. Amor, M. Daoudi, A. Srivastava, S. Berretti, Shape analysis of local facial patches for 3D facial expression recognition, Pattern Recognition 44 (8) (2011) 1581–1589.

[31] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: Proceedings of IEEE Computer Vision and Pattern Recognition, 2005.

[32] L. Yin, Z. Wei, Y. Sun, J. Wang, M. Rosato, A 3D facial expression database for facial behavior research, in: Proceedings of Seventh International Conference on Automatic Face and Gesture Recognition, 2006, pp. 211–216.

[33] X. Peng, M. Bennamoun, A. Mian, A training-free nose tip detection method from face range images, Pattern Recognition 44 (3) (2011) 544–558.

[34] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, Shape distributions, ACM Transactions on Graphics 21 (4) (2002) 807–832.

[35] C. Brown, Some mathematical and representational aspects of solid modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 3 (1981) 444–453.

[36] C. Hsu, C. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks 13 (2002) 415–425.

[37] X. Qi, Y. Han, Incooperating multiple SVMs for automatic image annotation, Pattern Recognition 40 (2006) 728–741.

[38] J. Platt, Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, MIT Press, Cambridge, MA, USA, 1999.

[39] V. Struc, N. Pavesic, Gabor-based kernel partial-least-squares discrimination features for face recognition, Informatica 20 (1) (2009) 115–138.

[40] V. Struc, N. Pavesic, The complete Gabor–Fisher classifier for robust face recognition, EURASIP Advances in Signal Processing 2010 (2010) 26.

[41] L. Chen, H. Liao, M. Ko, J. Lin, G. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.

[42] W. Zheng, L. Zhao, C. Zou, An efficient algorithm to solve the small sample size problem for LDA, Pattern Recognition 37 (5) (2004) 1077–1079.

[43] A. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition, IEEE Transactions on Circuits and Systems for Video Technology 14 (1) (2004) 4–20.

[44] F. Al-Osaimi, M. Bennamoun, A. Mian, Spatially optimized data-level fusion of texture and shape for face recognition, IEEE Transactions on Image Processing 21 (2) (2012) 859–872.

Yinjie Lei received his M.Sc. degree from Sichuan University, China in the area of image processing. He is currently working on his Ph.D. degree in Computer Science at the University of Western Australia. His research interests include image and text understanding, 3D face processing and recognition, machine learning and statistical pattern recognition.

Mohammed Bennamoun received his M.Sc. degree in control theory from Queen's University, Kingston, ON, Canada, and the Ph.D. degree in computer vision from Queensland University of Technology (QUT), Brisbane, Australia. He lectured Robotics at Queen's University and then joined QUT in 1993 as an Associate Lecturer. He is currently a Winthrop Professor and has been the Head of the School of Computer Science and Software Engineering, The University of Western Australia (UWA), Perth, Australia for five years (Feb, 2007–Feb, 2012). He was the Director of a university center at QUT, i.e., the Space Centre for Satellite Navigation, from 1998 to 2002. He was an Erasmus Mundus Scholar and a Visiting Professor at the University of Edinburgh, Edinburgh, U.K., in 2006. He was also a Visiting Professor at Centre National de la Recherche Scientique and Telecom Lille1, Villeneuve d'Ascq, France, in 2009, Helsinki University of Technology, Espoo, Finland, in 2006, and The University of Bourgogne, Dijon, France, and Paris 13, Villetaneuse, France, from 2002 to 2003. He is the co-author of the book "Object Recognition: Fundamentals and Case Studies", (Springer-Verlag, 2001) and the coauthor of an Edited book on "Ontology Learning and Knowledge Discovery Using the Web" published in 2011. He has published over 150 journal and conference publications and secured highly competitive national grants from the Australian Research Council (ARC). Some of these grants were in collaboration with industry partners (through the ARC Linkage Project scheme) to solve real research problems for industry, including Swimming Australia, the West Australian Institute of Sport, Beaulieu Pacific (a textile company), and AAM-GeoScan. He worked on research problems and collaborated (through joint publications, grants, and supervision of Ph.D. students) with researchers from different disciplines, including animal biology, speech processing, biomechanics, ophthalmology, dentistry, linguistics, robotics, and radiology. He collaborated with researchers from within Australia (e.g. CSIRO), as well as internationally (e.g. Germany, France, Finland, and USA). His areas of interest include control theory, robotics, obstacle avoidance, object recognition, artificial neural networks, signal/image processing, and computer vision (particularly 3D).

Amar A. El-Sallam received B.Sc. and M.Sc. in EEng from Assiut University in Egypt and a Ph.D. from Curtin University, Australia. He received several scholarships including the highly competitive and prestigious IPRS, and the Australian Telecommunication Cooperative Research Centre (ATcrc) and the Australian Telecommunication Research Institute (ATRI/WATRI) scholarships. From 2006 to 2009, Amar was a postdoctoral Research Fellow at the school of EEng at the University of Western Australia. Currently, he is a Research Assistant Professor at the School of Computer Science & Software Engineering at the same University. His current research includes; computer vision, biometrics, signal and image processing and wireless communications. Amar has over 42 publications, and 2 theses and was involved in three successful ARC grants. Recently, he and colleagues secured a grant from the Australian Institute of Sport and another from the Australian Ministry of health.