An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition

Ajmal S. Mian, Mohammed Bennamoun, and Robyn Owens

Abstract—We present a fully automatic face recognition algorithm and demonstrate its performance on the FRGC v2.0 data. Our algorithm is multimodal (2D and 3D) and performs hybrid (feature based and holistic) matching in order to achieve efficiency and robustness to facial expressions. The pose of a 3D face along with its texture is automatically corrected using a novel approach based on a single automatically detected point and the Hotelling transform. A novel 3D Spherical Face Representation (SFR) is used in conjunction with the Scale-Invariant Feature Transform (SIFT) descriptor to form a rejection classifier, which quickly eliminates a large number of candidate faces at an early stage for efficient recognition in case of large galleries. The remaining faces are then verified using a novel region-based matching approach, which is robust to facial expressions. This approach automatically segments the eyes-forehead and the nose regions, which are relatively less sensitive to expressions and matches them separately using a modified Iterative Closest Point (ICP) algorithm. The results of all the matching engines are fused at the metric level to achieve higher accuracy. We use the FRGC benchmark to compare our results to other algorithms that used the same database. Our multimodal hybrid algorithm performed better than others by achieving 99.74 percent and 98.31 percent verification rates at a 0.001 false acceptance rate (FAR) and identification rates of 99.02 percent and 95.37 percent for probes with a neutral and a nonneutral expression, respectively.

Index Terms—Biometrics, face recognition, rejection classifier, 3D shape representation.

1 Introduction

Biometrics are physiological (for example, fingerprints and face) and behavioral (for example, voice and gait) characteristics used to determine or verify an individual's identity [6]. Verification is performed by matching an individual's biometric with the template of the claimed identity only. Identification, on the other hand, is performed by matching an individual's biometric with the template of every identity in the database.

The human face is an easily collectible, universal, and nonintrusive biometric [19], which makes it ideal for applications in scenarios where fingerprinting or iris scanning are impractical (for example, surveillance) or undesirable due to problems of social acceptance [20]. However, face recognition is a challenging problem because of the diversity in faces and variations caused by expressions, gender, pose, illumination, and makeup. Considerable work has been done in this area resulting in a number of face recognition algorithms [50]. These algorithms are categorized from two different perspectives, namely, the type of data and the type of approach they use. From the first perspective, face recognition algorithms are divided into 1) 2D face recognition (which use 2D gray-scale or color images), 2) 3D face recognition (which use 3D range images or pointclouds of faces), and 3) multimodal face recognition algorithms (which use both 2D and 3D facial data), for example, [28].

Appearance-based (2D) face recognition algorithms were the first to be investigated due to the wide spread availability of low-cost cameras. However, 2D face recognition is sensitive to illumination, pose variations, facial expressions [50], and makeup. A comprehensive survey of 2D face recognition algorithms is given by Zhao et al. [50]. They also categorize face recognition from the second perspective into 1) holistic, 2) feature based (referred to as region based[1] in this paper), and 3) hybrid-matching face recognition algorithms. Holistic algorithms match the faces as a whole for recognition. Examples of this method include the eigenfaces of Turk and Pentland [44] that use the Principal Component Analysis (PCA), the Fisherfaces [4] that use Linear Discriminant Analysis (LDA), methods based on the Independent Component Analysis (ICA) [3], Bayesian methods [37], and Support Vector Machine (SVM) methods [40]. Neural networks [25] have also been used for holistic face recognition. The region-based methods extract regions like the eyes, nose, and mouth and then match these for face recognition. These methods are based on the distances/angles between facial regions or their appearances. Examples of this category include [14], [43]. The graph-matching approach [47] is one of the most successful region-based approaches [50]. Region-based methods can prove useful in case of variations (for example, illumination and expression) in the images [50]. Hybrid methods use a combination of the holistic and feature-based matching for improved recognition performance. An example of hybrid method combines the eigenfaces, eigeneyes, and eigennose [39]. Other examples include the flexible appearance model-based method in [23] and [17].

A detailed survey of 3D and multimodal face recognition is given by Bowyer et al. [7]; however, a brief survey is included here for completeness. Chua et al. [13] extracted point

• The authors are with the School of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway, Crawley, Western Australia, 6009.
E-mail: {ajmal, bennamou}@csse.uwa.edu.au, robyn.owens@uwa.edu.au.

Manuscript received 24 May 2006; revised 2 Nov. 2006; accepted 9 Jan. 2007; published online 1 Feb. 2007.
Recommended for acceptance by S. Baker.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0398-0506.
Digital Object Identifier no. 10.1109/TPAMI.2007.1105.

[1] We call it region based to differentiate it from features that are extracted by feature extraction algorithms.

signatures [12] of the *rigid parts* of the face for expression-invariant face recognition. They reported 100 percent recognition results but on a small gallery of six subjects. Xu et al. [48] performed automatic 3D face recognition by combining global geometric features with local shape variation and reported 96.1 percent and 72.6 percent recognition rates when using a gallery of 30 and 120 subjects, respectively. Notice the 23.5 percent drop in recognition rate when the gallery size is increased four times. Medioni and Waupotitsch [30] also used a variant to the Iterative Closest Point (ICP) algorithm for 3D face recognition and reported a recognition rate of 98 percent on a gallery of 100 subjects. The above results have been achieved using very small databases and their scalability to large databases is highly questionable.

Fortunately, the FRGC v2.0 data is now publicly available and results achieved on this database are more compelling as there are more subjects and greater quantity of images in this database (see Section 2 for details). An example of results achieved on the FRGC v2.0 database is the Adaptive Rigid Multiregion Selection (ARMS) approach of Chang et al. [11] who report a recognition rate of 92 percent. Another example is the annotated deformable model approach of Passalis et al. [38] who achieved an average verification rate of 85.1 percent at a 0.001 false acceptance rate (FAR) on the FRGC v2.0 data. The performance of both these approaches is significantly affected by facial expressions. The rank-one recognition rate of ARMS [11] drops from approximately 98 percent to 88 percent as a result of nonneutral facial expressions. Likewise, the verification rate at 0.001 FAR of the deformable model approach [38] drops from approximately 94.9 percent to 79.4 percent in the presence of nonneutral expressions. For a summary of more results achieved on the FRGC v2.0 database, the interested reader is referred to [42].

Existing approaches to multimodal face recognition generally perform separate matching on the basis of 2D and 3D faces and then fuse the results at the metric level. Chang et al. [9] used a PCA-based approach for separate 2D and 3D face recognition and fused the matching scores. They reported a recognition rate of 93 percent and 99 percent for 3D and multimodal face recognition, respectively, using a gallery of 275 subjects. Wang et al. [46] used Gabor Wavelet filters in the 2D domain and the point signatures [12] in the 3D domain and fused the results using SVM. They reported a recognition rate of above 90 percent on a gallery of 50 subjects.

Bronstein et al. [8] proposed an expression-invariant multimodal face recognition algorithm. They assume the 3D facial surface to be isometric and remove the effects of expressions by finding an isometry-invariant representation of the face. The downside of this approach is that it also attenuates some important discriminating features like the 3D shape of the nose and the eye sockets. They used a database of only 30 subjects and did not discuss how an open mouth expression is handled by their algorithm.

Lu et al. [28] used feature detection and registration with the ICP [5] algorithm in the 3D domain and LDA in the 2D domain for multimodal face recognition. They handle pose variations by matching partial scans of the face to complete face models. The gallery 3D faces are also used to synthesize novel appearances (2D faces) with pose and illumination variations in order to achieve robustness during 2D face matching. They achieved a multimodal recognition rate of 99 percent for neutral faces and 77 percent recognition rate for smiling faces using a database of 200 gallery and 598 probe faces. The performance on neutral-expression faces is quite impressive. However, the recognition rate drops by 22 percent in the case of smiling faces. Besides this drop in the recognition rate, there are two points to be noted here. First, the database used is smaller than the FRGC v2.0. Second, only smiling faces are tested. It would be interesting to see how this system performs on the FRGC v2.0 data, which is not only larger but also contains more challenging nonneutral facial expressions like blown cheeks and open mouth.

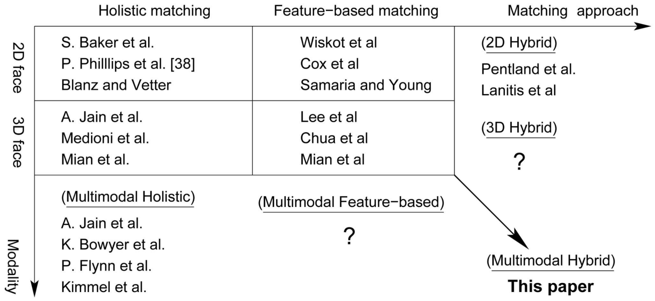


Fig. 1. Illustration of current research in the area of face recognition. This figure contains only a few examples for illustration and is far from being exhaustive. The areas where there is a lack of adequate research are marked by a "?".

Examples of multimodal face recognition algorithms that have been tested on the FRGC v2.0 database include Maurer et al. [29] who measure the performance of the Geometrix ActiveID [15] on the FRGC v2.0 data and achieved a multimodal verification rate of 99.2 percent at 0.001 FAR for faces with a neutral expression. Maurer et al. [29] do not report separate results for all nonneutral-expression faces; however, their combined results (neutral versus all) show a drop of 3.4 percent in the verification rate. Huskën et al. [18] use a hierarchical graph-matching approach and achieve a verification rate of 96 percent at 0.001 FAR for neutral versus all faces using the FRGC v2.0 data. However, they do not report separate results for faces with nonneutral expressions. The survey of Bowyer et al. [7] concludes that there is still a need for *improved sensors, recognition algorithms, and experimental methodology*.

Bowyer et al. [7] state that multimodal face recognition outperforms both 2D and 3D face recognition alone. Zhao et al. [50] argue that the hybrid-matching methods "could potentially offer the best of the two types of methods" (that is, feature-based and holistic matching). In this paper, we combine these two thoughts and present a fully automatic multimodal hybrid face recognition algorithm, which requires no user intervention at any stage. Using the FRGC v2.0 data (which is the largest available of its kind), we demonstrate that by exploiting multimodal hybrid-matching techniques, very high face recognition performance can be achieved both in terms of efficiency and accuracy.

Fig. 1 illustrates the current research in the area of face recognition by plotting the different categories along orthogonal axes. The various matching approaches are plotted along the x-axis, whereas different modalities are plotted along the y-axis. Another dimension could be the temporal one, that is, recognition from single images or video sequences. Video-based recognition has mainly been performed on 2D holistic faces. However, it is not discussed here as it is outside the scope of this paper. Examples of some research groups working along each dimension are also given in Fig. 1, and the areas where there is a lack of adequate research are marked by a "?". Note that Fig. 1 gives only a few examples of each category and is far from being exhaustive. In Fig. 1, it is clear that there is a lack of research in the area of 3D hybrid and multimodal feature-based face recognition. This paper covers both these areas and presents a multimodal hybrid face recognition algorithm. This paper also addresses two major problems in 3D face recognition. The first problem is of facial expressions. Although 3D face recognition has the potential to achieve higher accuracy, it is more sensitive to facial expressions compared to its 2D counterpart. The second problem addressed in this paper is of computational efficiency. Three-dimensional face recognition is computationally expensive, and a brute force matching approach does not scale well to large galleries such as the FRGC v2.0 (Section 2).

An advantage of 3D data is that it can be used to correct the pose of both the 3D and its corresponding 2D face, which is the *first contribution* of our paper. We present a fully automatic algorithm for the pose correction of a 3D face, and its corresponding 2D colored image. Existing techniques typically require the manual identification of multiple landmarks on a face for pose correction (for example, [10]). Our approach is based on the automatic detection of a single point (the nose tip) on the face. It then iteratively corrects the pose using the Hotelling transform [16]. The pose correction measured from the 3D face is also used to correct the *3D pose* of its corresponding 2D face.

The *second contribution* of our paper (which is a major one) is a novel holistic 3D Spherical Face Representation (SFR). SFR is efficiently calculated and used in conjunction with the Scale-Invariant Feature Transform (SIFT) descriptor [26] to form a rejection classifier, which quickly eliminates a large number of ineligible candidate faces from the gallery at an early stage. SFR is a low-cost global 3D face descriptor. SIFTs are 2D local descriptors and have been successfully used for object recognition under occlusions. In this paper, the utility of SIFT for face recognition under illumination and expression variations has been explored using the FRGC v2.0 database (Section 2).

Recently, Lin and Tang [24] proposed a SIFT-Activated Pictorial Structure (SAPS) and combined it with three other classifiers for face recognition. There are three major differences of our work from SAPS. First, SAPS requires more than one training face, whereas we use only one face. Second, SAPS computes SIFTs at only those keypoints that contain irregular details of a face, whereas we compute SIFTs at all keypoints (see Section 4.2). Third, the individual performance of SAPS classifier is not reported in [24], whereas we report the individual performance of our SIFT classifier in Fig. 15.

After rejection, the remaining faces are verified using a novel region-based matching [32], [33] approach, which is our *third major contribution*. Our region-based matching approach is robust to facial expressions as it automatically segments those regions of the face that are relatively less sensitive to expressions, that is, the eyes-forehead and the nose. These regions are separately matched using a *modified* ICP algorithm, which exploits the dissimilarities between faces [32]. The motivation behind our region-based 3D matching approach comes from three important findings in the 2D face recognition survey by Zhao et al. [50]. One, that the upper part of the face is more significant for recognition compared to the lower part. Two, that region-based matching can prove useful in the case of expression and illumination variations. Three, that the eyes, the forehead, and the nose are less sensitive to facial expressions compared to the mouth and the cheeks. Our approach simply extends these ideas to the 3D face recognition case. Note that prior attempts have been made to recognize 3D faces using regions or segments. However, a curvature-based segmentation was used in those cases (see [7] for details) as opposed to the region-based segmentation in our case. Moreover, the component-based face recognition proposed by Huang et al. [17] performs recognition on the basis of 2D components. Our 3D segmentation is different from the 2D case because the segmented regions need not be rectangular and can also vary in size. Moreover, a pixel-to-pixel correspondence or 100 percent overlap is not required between the segmented regions for matching.

Each matching engine (SFR-SIFT, eyes-forehead, and nose) results in a similarity matrix. The similarity matrices are normalized and fused using a confidence weighted summation rule to achieve better recognition performance. This results in a single similarity matrix that is used to compile the identification and verification performance of our algorithm.

This paper is an extension of our work presented in [32], [33], [34] and is organized as follows. In Section 2, we give some information regarding the FRGC v2.0 data and Experiment 3. Section 3 explains our automatic 3D face detection and normalization algorithm along with qualitative and quantitative results. Section 4 gives details of our novel SFR. It also explains the SFR-SIFT-based rejection classifier along with results on the FRGC v2.0 data. Section 5 gives details of our automatic face segmentation and region-based matching algorithm. Section 6 lists and compares the recognition results of our novel multimodal hybrid algorithm with and without using the rejection classifier. In Section 7, we give the limitations of our pose correction and recognition algorithms and give directions for improvements. Finally, in Section 8, we conclude our findings.

2 The FRGC v2.0 Description

We performed our experiments on the FRGC version 2.0 [41] data set.[2] The FRGC lists a number of experiments; however, we focus on Experiment 3, that is, matching 3D faces (shape and texture) to 3D faces (shape and texture) in this paper. The FRGC v2.0 data for Experiment 3 comprise 4,950 3D faces along with their corresponding texture maps acquired with the Minolta Vivid scanner [36]. The spatial resolution of the scanner is 480×640. However, the resolution of faces in the database varies because they were scanned at different distances from the scanner and possibly using different lenses. The data consists of frontal views of subjects mostly acquired from the shoulder level up. The subjects include males and females aged 18 years and above. Some of the subjects have facial hair, but none of them is wearing glasses. There are minor pose variations and major illumination, as well as expression variations, in the database. More detailed statics on the database are given by Phillips et al. [41].

The 3D faces (shape) in the FRGC data are available in the form of four matrices, each of size 480×640. The first matrix is a binary mask indicating the valid pixels (or points) in the remaining three matrices that respectively contain the x, y, and z-coordinates of the pixels. The 2D faces (texture maps) are 480×640 color images having a one-to-one correspondence to their respective 3D face. The texture maps are

[2] This work was not submitted to the FRGC workshop held in March 2006 as it was not ready then.

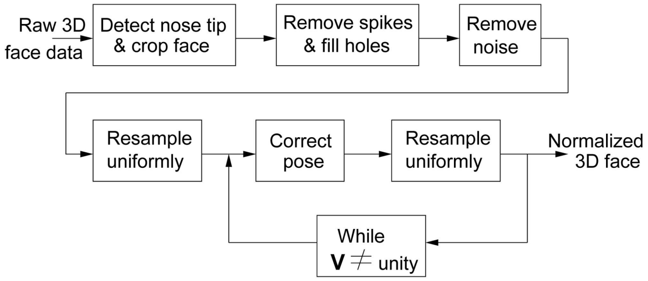


Fig. 2. Block diagram of the face normalization. $\mathbf{V}$ is a rotation matrix given in (4).

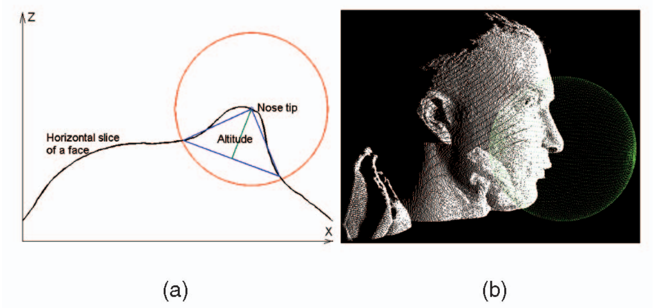


Fig. 3. (a) Nose tip detection. (b) A sphere centered at the nose tip of a 3D pointcloud of a face is used to crop the face.

correctly registered to the 3D faces in most cases, however, some examples of incorrect registration can be found in the database. The 3D faces are noisy and contain spikes, as well as holes (Fig. 4). In this paper, we represent a 3D face as a three-dimensional vector $[x_i, y_i, z_i]^\top$ of the x, y, and z-coordinates of the pointcloud of a face ($i = 1 \dots n$, where n is the number of points). For the purpose of pose correction, we represent a 2D face as a five-dimensional vector $[u_i, v_i, R_i, G_i, B_i]^\top$, where u and v are the pixel coordinates, and R, G, and B are their corresponding red, green, and blue components. Since the 3D and 2D faces are registered in most cases, the pixel coordinates u and v of the 2D face can be replaced with the absolute coordinates x and y of its corresponding 3D face. For the purpose of matching, we represent the 2D face as a gray-scale image.

Finally, the data is divided into three sets (based on acquisition time), namely, Spring2003, Fall2003, and Spring2004. The FRGC explicitly specifies that Spring2003 be used for training and the remaining two sets be used for validation. Although our algorithms do not have a training phase like the PCA, we used the training data for setting our thresholds and for tuning. The validation set contains 4,007 3D faces along with their texture maps. The number of subjects in the validation set is 466. We chose one 3D face along with its texture to make a gallery of 466 individuals for our identification experiments. The FRGC also gives a target and a query set of 4,007 images each. According to the FRGC protocol, each query face must be matched with each target face, which amounts to 16 million similarity scores. The results reported in Section 6 are obtained using the validation set only so that they are compatible with the FRGC.

3 3D and 2D Face Normalization

Fig. 2 shows the block diagram of our automatic 3D and 2D face normalization algorithm. Details of the different components of the block diagram are given below.

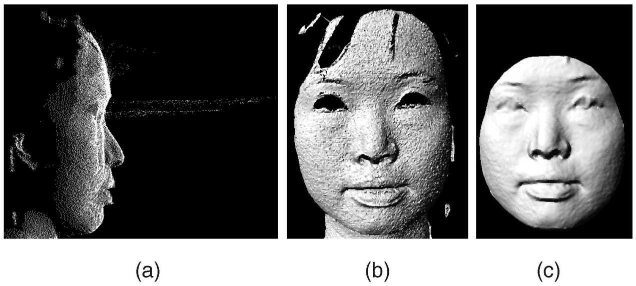


Fig. 4. (a) A pointcloud of a face shows spikes. (b) A shaded view of the same face shows noise. Spike removal has also resulted in holes. (c) Shaded view of the face after complete preprocessing (that is, cropping, hole filling, denoising, and resampling).

3.1 Face Localization and Denoising

Since the FRGC data contains faces mostly acquired from the shoulder level up, an important preprocessing step was to localize the face. Since processing 3D data is computationally expensive, we detect the nose tip in the first step in order to crop out the required facial area from the 3D face for further processing. The nose tip is detected using a coarse to fine approach as follows: Each 3D face is horizontally sliced (Fig. 3a) at multiple steps d_v. Initially, a large value is selected for d_v to improve speed, and once the nose is coarsely located, the search is repeated in the neighboring region with a smaller value of d_v. The data points of each slice are interpolated at uniform intervals to fill in any holes. Next, circles centered at multiple horizontal intervals d_h on the slice are used to select a segment from the slice, and a triangle is inscribed using the center of the circle and the points of intersection of the slice with the circle, as shown in Fig. 3a. Once again, a coarse to fine approach is used for selecting the value of d_h for performance reasons. The point that has the maximum altitude triangle associated with it is considered to be a potential nose tip on the slice and is assigned a confidence value equal to the altitude. This process is repeated for all slices resulting in one candidate point per slice along with its confidence value. These candidate points correspond to the nose ridge and should form a line in the xy-plane. Some of these points may not correspond to the nose ridge. These are outliers and are removed by robustly fitting a line to the candidate points using Random Sample Consensus (RANSAC) [22]. Out of the remaining points, the one that has the maximum confidence is taken as the nose tip, and the above process is repeated at smaller values of d_v and d_h in the neighboring region of the nose tip for a more accurate localization.

A sphere of radius r centered at the nose tip is then used to crop the 3D face (see Fig. 3) and its corresponding registered 2D face. A constant value of $r = 80\ mm$ was selected in our experiments. This process crops an elliptical region (when viewed in the xy-plane) from the face with vertical major axis and horizontal minor axis. The aspect ratio (major axis to minor axis ratio) of the ellipse varies with the curvature of the face. For example, the narrower a face is, the greater is its aspect ratio. Fig. 5a shows a histogram of the aspect ratios of 466 different faces. Once the face is cropped, the outlier points causing spikes (see Fig. 4a) in the 3D face are removed. We defined outlier points as the ones whose distance is greater than a threshold d_t from any one of its 8-connected neighbors. d_t is automatically calculated using $d_t = \mu + 0.6\sigma$ (where μ is the mean distance between neighboring points, and σ is its standard deviation). After removing spikes, the 3D face and

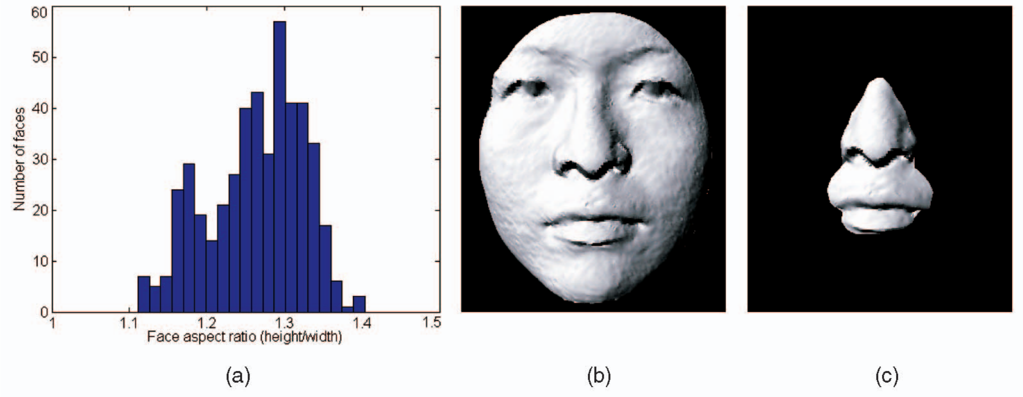


Fig. 5. (a) Histogram of the aspect ratios of 466 individuals. (b) A shaded view of a sample 3D face with low aspect ratio. (c) The aspect ratio considerably increases when a smaller region is cropped from the face using a combination of radius and depth thresholds.

its corresponding 2D face are resampled on a uniform square grid at $1\ mm$ resolution. The removal of spikes may result in holes (see Fig. 4b) in the 3D face, which are filled using cubic interpolation. Resampling the 2D face on a similar grid as the 3D face ensures that a one-to-one correspondence is maintained between the two. Since noise in 3D data generally occurs along the viewing direction (z-axis) of the sensor, the z-component of the 3D face (range image) is denoised using median filtering (see Fig. 4c).

3.2 Pose Correction and Resampling

Once the face is cropped and denoised, its pose is corrected using the Hotelling transform [16], which is also known as the Principle Component Analysis (PCA). Let $\mathbf{P}$ be a $3 \times n$ matrix of the x, y, and z-coordinates of the pointcloud of a face (1)

$$\mathbf{P} = \begin{bmatrix} x_1\ x_2\ \ldots\ x_n \\ y_1\ y_2\ \ldots\ y_n \\ z_1\ z_2\ \ldots\ z_n \end{bmatrix}. \tag{1}$$

The mean vector $\mathbf{m}$ and the covariance matrix $\mathbf{C}$ of $\mathbf{P}$ are given by

$$\mathbf{m} = \frac{1}{n}\sum_{k=1}^{n} P_k, \quad \text{and} \tag{2}$$

$$\mathbf{C} = \frac{1}{n}\sum_{k=1}^{n} P_k P_k^T - \mathbf{m}\mathbf{m}^T, \tag{3}$$

where P_k is the kth column of $\mathbf{P}$. Performing PCA on the covariance matrix $\mathbf{C}$ gives us a matrix $\mathbf{V}$ of eigenvectors and a diagonal matrix $\mathbf{D}$ of eigenvalues such that

$$\mathbf{CV} = \mathbf{DV}. \tag{4}$$

$\mathbf{V}$ is also a rotation matrix that aligns the pointcloud $\mathbf{P}$ on its principal axes, that is, $\mathbf{P}' = \mathbf{V}(\mathbf{P} - \mathbf{m})$.

Pose correction may expose some regions of the face (especially around the nose), which are not visible to the 3D scanner. These regions have holes that are interpolated using cubic interpolation. The face is resampled once again on a uniform square grid at $1\ mm$ resolution and the above process of pose correction and resampling is repeated until $\mathbf{V}$ converges to an identity matrix (see block diagram in Fig. 2). The faces with small aspect ratio (Fig. 5b) are prone to misalignment errors along the z-axis. Therefore, after pose correction along the x and y-axes, a smaller region is cropped from the face using a radius of $50\ mm$ (centered at the nose tip) and a depth threshold equal to the mean depth of the face (with $r = 80\ mm$). This results in a region with a considerably higher aspect ratio (Fig. 5c), which is used to correct the facial pose along the z-axis.

Resampling the faces on a uniform square grid has another advantage that all the faces end up with equal resolution. This is very important for the accuracy of the 3D matching algorithm (Section 5.1), which is based on measuring point-to-point distances. Difference in the resolution of the faces can bias the similarity scores in favor of faces that are more densely sampled. This makes sense because, for a given point in a probe face, there are more chances of finding a closer point in a densely sampled gallery face compared to a rarely sampled one.

$\mathbf{V}$ is also used to correct the *3D pose* of the 2D face corresponding to the 3D face. The R, G, and B pixels are mapped onto the pointcloud of the 3D face and rotated using $\mathbf{V}$. This may also result in missing pixels, which are interpolated using cubic interpolation. To maintain a one-to-one correspondence with the 3D face, as well as for scale normalization, the 2D colored image of the face is also resampled in exactly the same manner as the 3D face. Fig. 6 shows a sample face (2D and 3D) before and after normalization. It is important to note that this scale normalization of the 2D face is different from the one found in the existing literature. Previous methods (for example, [10]) are based on *manually* identifying two points on the face (generally, the corners of the eyes) and normalizing their distance to a prespecified number of pixels. As a result, the distance (measured in pixels) between the eyes of all individuals end up the same irrespective of the absolute distance. This brings the faces closer in face feature space, hence making classification more challenging. On the other hand, with our 3D-based normalization algorithm, the distance between the eyes of each individual may be different as it is a function of the absolute distance between the eyes. Thus, the faces remain comparatively far in face space, which results in a more accurate classification.

3.3 Pose Correction Results

Fig. 7 shows some sample 3D and their corresponding 2D faces from the FRGC v2.0 data set after pose correction. A

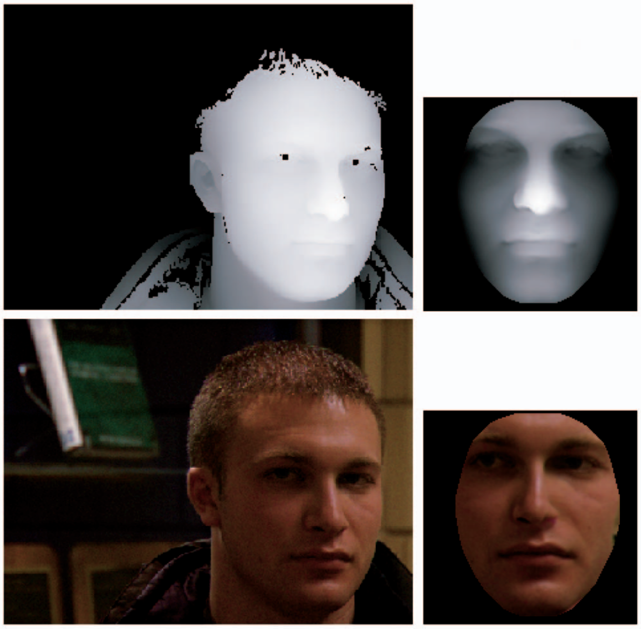


Fig. 6. A 3D face and its corresponding 2D face (colored) before and after pose correction and normalization.

qualitative analysis of these results shows that our algorithm is robust to facial expressions and hair that covers the face. For quantitative analysis, the pose of each face must be compared with some ground truth. Since ground truth was not available, we pairwise registered the 3D faces belonging to the same identities with each other (all possible combinations C_2^n, where n is the number of 3D faces belonging to the same identity) using the ICP [5] algorithm. The translation and rotation errors between these faces are presented in Figs. 8 and 9, respectively. The maximum absolute mean translation and rotation errors between the faces were $0.48\,mm$ and 0.99 degrees, respectively. Note that the translation error is the error in nose-tip localization.

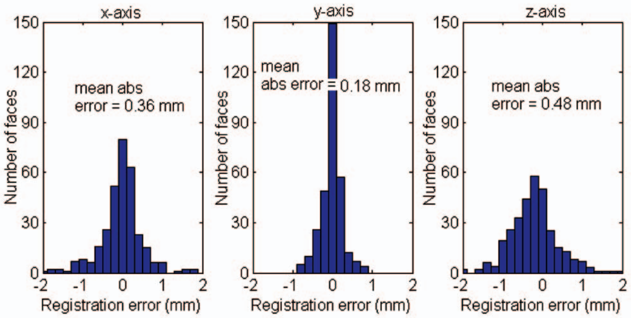


Fig. 8. Translation errors between the 3D faces of the same identities after automatic pose correction. The errors are less than $0.5mm$, which is half the spatial resolution of the 3D faces.

4 A Low-Cost Rejection Classifier

A rejection classifier is defined as one that quickly eliminates a large percentage of the candidate classes with high probability [2]. A rejection classifier is "an algorithm ψ that given an input, $x \epsilon S$, returns a set of class labels $\psi(x)$ such that $x \epsilon W_i \Rightarrow i \epsilon \psi(x)$" [2], where x is a measurement vector, $S = \Re^d$ is a classification space of d measurements, and W_i is the ith class such that $W_i \subseteq S$. The effectiveness $\text{Eff}(\psi)$ of a rejection classifier is the expected cardinality of the rejector output $E_{x\epsilon S}(|\,\psi(x)\,|)$ divided by the total number of classes M (5) [2]

$$\text{Eff}(\psi) = \frac{E_{x\epsilon S}(|\,\psi(x)\,|)}{M}. \tag{5}$$

In our case, M is the size of the gallery. The smaller the value of $\text{Eff}(\psi)$, the better is the rejection classifier. A rejection classifier is necessary to perform efficient recognition when using large databases. For example, in our case, there were 3,541 probes and 466 faces in the gallery. A brute force matching approach would have required $466 \times 3,451 = 16,08,166$ comparisons. A rejection classifier of $\text{Eff}(\psi) = 0.03$ would reduce this to only 48,245 comparisons. Fig. 10

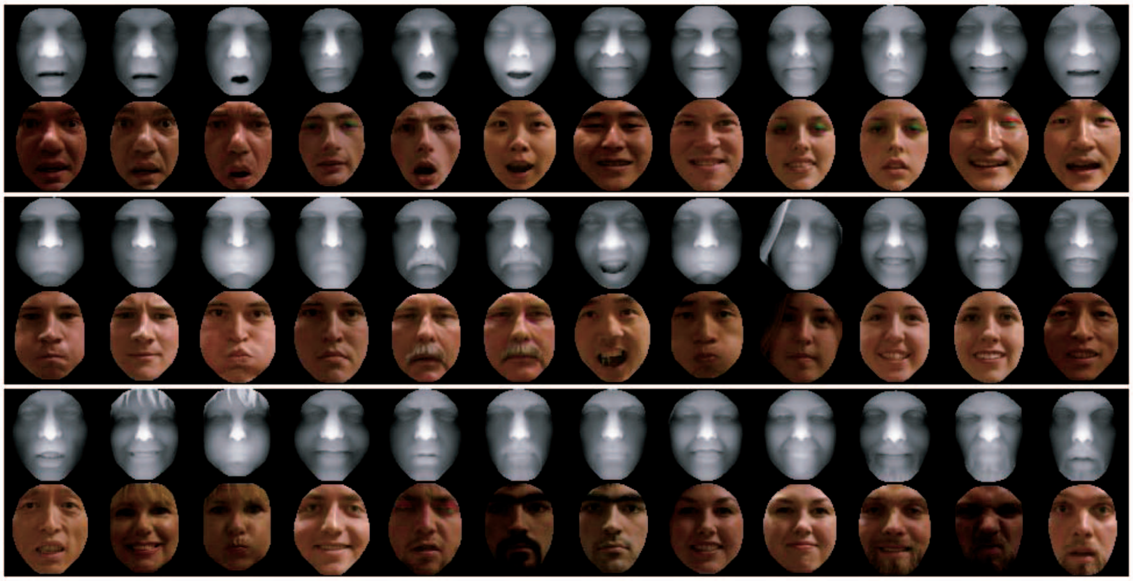


Fig. 7. Sample 3D faces and their corresponding 2D (colored) faces after pose correction and normalization. All these faces were also correctly recognized using our algorithm explained in Section 5 (even though some of them had hair) except for the second last face in the last column.

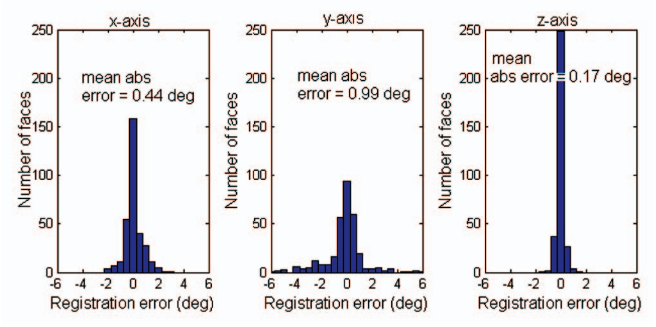


Fig. 9. Rotation errors between the 3D faces of the same identities after automatic pose correction. All errors are less than 1 degree.

shows the block diagram of our complete 3D face recognition algorithm including the rejection classifier.

4.1 Spherical Face Representation (SFR)

We present a novel SFR and compare it to two existing 3D representations, that is, the spin images [21] and the tensor representation [31], [35]. A spin image is generated by spinning an image (for example, of size 6×6 in Fig. 11b) around the normal of a point (the nose tip in our case) and summing the face points as they pass through the bins of the image. The tensor representation [31], [35] is the quantization of the surface area of a face into a 3D grid. Fig. 11c shows a $10 \times 10 \times 10$ tensor over a facial pointcloud.

Intuitively, an SFR can be imagined as the quantization of the pointcloud of a face into spherical bins centered at the nose tip. Fig. 11a graphically illustrates an SFR of three bins. To compute an n bin SFR, the distance of all points from the origin is calculated. These distances are then quantized into a histogram of $n+1$ bins. The outermost bin is then discarded since it is prone to errors (for example, due to hair). An SFR can be efficiently computed by exploiting Matlab functionality. In our experiments, we used a 15 bin SFR. Fig. 12a shows eight SFRs each of two individuals (under a neutral expression) plotted on the y-axis. The SFRs belonging to the same individual follow a similar curve shape, which is different from that of a different identity. Fig. 12b shows the SFR variation of an identity under nonneutral expressions. The bold line represents the SFR under a neutral expression, whereas the thin ones represent the SFRs under a nonneutral expression. The similarity between a probe and gallery face is computed by measuring the pointwise euclidean distance between their SFRs.

The tensor representation has higher discriminating capability [31] compared to a spin image. The recognition

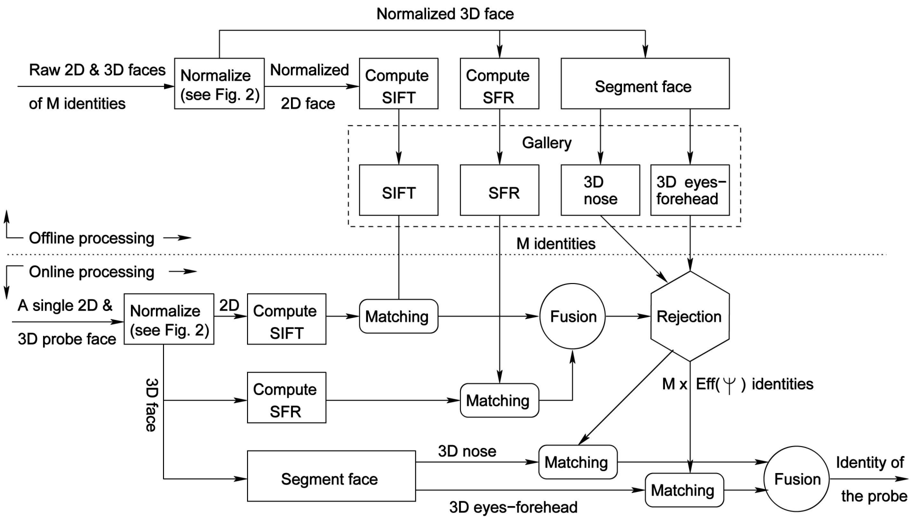


Fig. 10. Block diagram of our recognition algorithm (MMH_e). The dotted line separates the online and offline phases, whereas the dashed line shows the representations included in the gallery.

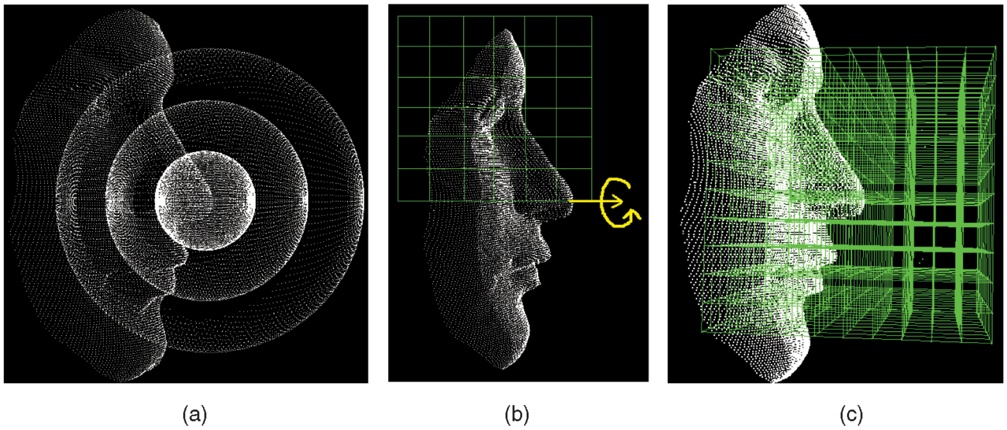
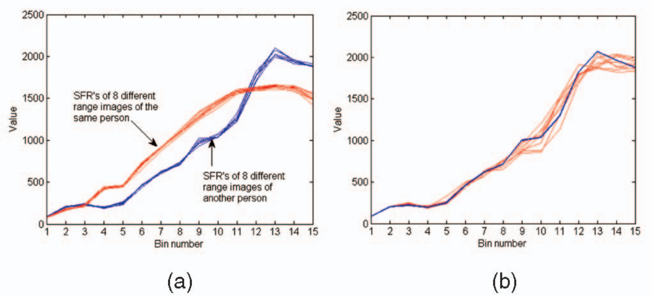

(a) (b) (c)

Fig. 11. Illustration of the (a) SFR, (b) spin image, and (c) tensor representation.



(a) (b)

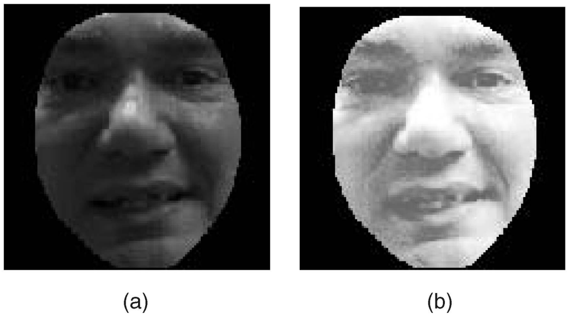
Fig. 12. (a) SFRs of two individuals (under a neutral expression) plotted on the y-axis (in different shades). Notice that the SFRs belonging to the same identity are quite similar, whereas those of different identities are dissimilar. (b) SFR variation of an identity under nonneutral expressions. The bold line represents the SFR under a neutral expression.

performance of a representation is directly related to its descriptiveness [35]. However, on the downside, higher descriptiveness of a representation makes it more sensitive to nonrigid deformations. Therefore, as a consequence of its higher descriptiveness, the tensor representation is more sensitive to facial expressions and is therefore not considered here. The descriptiveness of the SFR intuitively appears to be the lowest of the three, which should make it less sensitive to facial expressions (see Fig. 12b). Moreover, in terms of computational complexity, the SFR is the most efficient. Therefore, for use as a rejector, the SFR appears to be the best choice. A brief comparison of the SFR and the spin images is given in Section 4.3, whereas a more detailed experimental comparison can be found in our earlier work [34].

4.2 SIFT-Based Matching Engine

The SIFTs [26] are local 2D features calculated at keypoint locations. The interested reader is referred to Lowe's paper [26] for the details of the keypoint localization and the SIFT feature extraction. A brief description is provided here for completeness. A cascaded filtering approach (keeping the most expensive operation to the last) is used to efficiently locate the keypoints, which are stable over scale space. First, stable keypoint locations in scale space are detected as the scale space extrema in the Difference-of-Gaussian function convolved with the image. A threshold is then applied to eliminate keypoints with low contrast followed by the elimination of keypoints, which are poorly localized along an edge. Finally, a threshold on the ratio of principal curvatures is used to select the final set of stable keypoints. For each keypoint, the gradient orientations in its local neighborhood are weighted by their corresponding gradient magnitudes and by a Gaussian-weighted circular window and put in a histogram. Dominant gradient directions, that is, peaks in the histogram, are used to assign one or more orientations to the keypoint.

At every orientation of a keypoint, a feature (SIFT) is extracted from the gradients in its local neighborhood. The coordinates of the feature and the gradient orientations are rotated relative to the keypoint orientation to achieve orientation invariance. The gradient magnitudes are weighted by a Gaussian function giving more weight to closer points. Next, 4×4 sample regions are used to create orientation histograms, each with eight orientation bins forming a $4 \times 4 \times 8 = 128$ element feature vector. To achieve robustness to illumination changes, the feature vector is normalized to unit length, large gradient magnitudes are then thresholded so that they do not exceed 0.2 each, and the vector is renormalized to unit length.



(a) (b)

Fig. 13. A sample gray-scale 2D image of a face before (a) and after (b) histogram equalization.

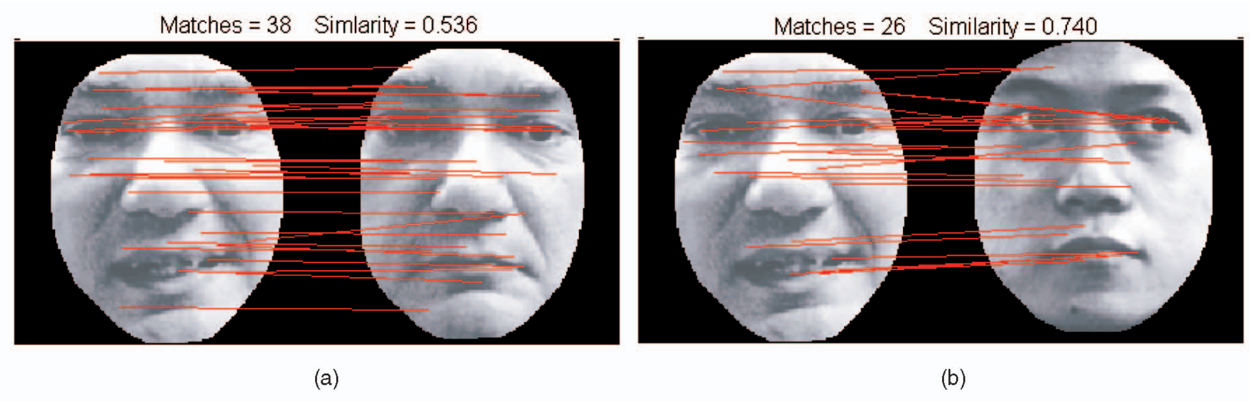
SIFTs have been successfully used for object recognition under occlusions, which is an intraclass recognition problem. We explore their use for face recognition using the FRGC v2.0 data (Experiment 3 texture only), which is challenging because it is an interclass recognition problem, and there are extensive illumination and expression variations in this data. Since SIFT is a local feature, we believe that it will be more robust to these variations compared to the PCA baseline performance. The SIFT descriptors were extracted from the 2D texture maps of faces after normalization with respect to pose and scale as described in Section 3.2. It is possible to normalize the illumination of the 2D faces using their corresponding 3D faces; however, this is not the focus of our paper. Therefore, we converted the colored images to gray scale and performed histogram equalization for illumination normalization. Fig. 13 shows the effect of histogram equalization on a sample face. The SIFT descriptors for these faces were then computed using Lowe's code [27]. The number of descriptors varied for each face with an average of 80 descriptors per image.

To calculate the similarity between a gallery and probe face, their SIFT descriptors were matched using the euclidean distance. Since the faces were registered, only those descriptors were considered, which were closely located. Moreover, only one-to-one matches were established, that is, a SIFT from the gallery was allowed to be a match to only one SIFT from the probe (see Fig. 14). The similarity score was taken as the mean euclidean distance between the best matching pairs of descriptors.

4.3 Rejection Classification Results

In our earlier work [34], we quantitatively compared the performance of the SFR (15 bins) to the spin images (size 15×15) [21] when used as rejection classifiers. For probes with a neutral expression, the spin images performed slightly better, whereas for probes with a nonneutral expression, the SFR performed better. This supported our argument that representations with lower descriptiveness are less sensitive to facial expressions, as discussed in Section 4.1. However, the SFR-based classifier is computationally much more efficient than the spin image classifier. A Matlab implementation on a 2.3-GHz Pentium IV machine took 6.8 ms to construct an SFR of a probe, match it with the 557 SFRs in the gallery, and reject a subset of the gallery, whereas the spin images took 2,863 ms for the same purpose.

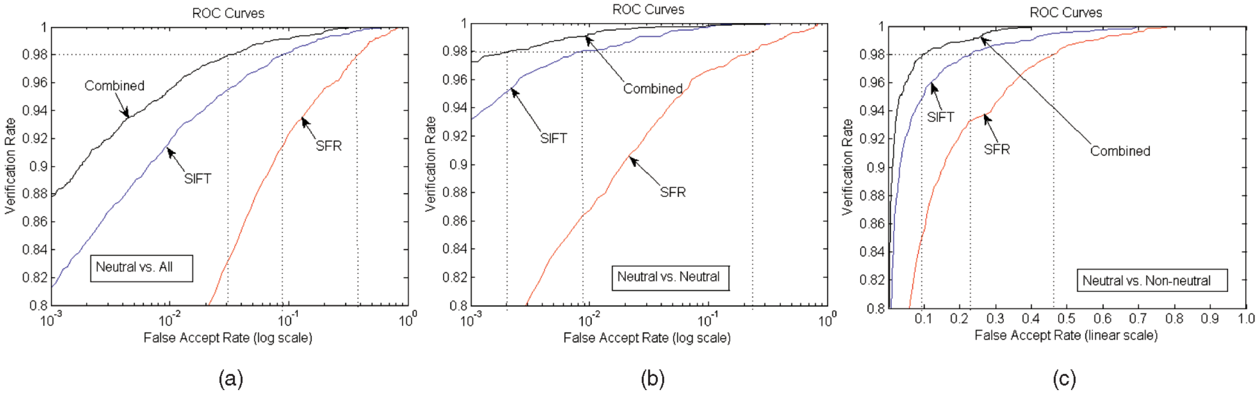
In this paper, we fused (see Section 5.2 for details) the similarity scores of the SFR and the SIFT descriptors using a


Matches = 38 Simlarity = 0.536
Matches = 26 Simlarity = 0.740

(a)

(b)

Fig. 14. SIFT matches between probe and gallery faces belonging to (a) the same identity and (b) different identities (lower value of similarity means a better match). The number and quality of matches are better in the case of (a).



(a)

(b)

(c)

Fig. 15. Rejection classification results. At the 98 percent verification rate, (a) $\mathrm{Eff}(\psi) = 0.036$ for all faces, (b) $\mathrm{Eff}(\psi) = 0.002$ for neutral versus neutral faces, and (c) $\mathrm{Eff}(\psi) = 0.09$ for neutral versus nonneutral faces.

weighted sum rule to achieve better rejection results (see Fig. 15). At the 98 percent verification rate, the effectiveness of our SFR-SIFT-based rejection classifier as per (5) is 0.036 for the entire database (that is, neutral versus all). Put another way, our rejection classifier will eliminate 97 percent of the gallery faces leaving only 3.6 percent to be verified at a later stage (out of 466, only 17 faces will be left for verification). A Matlab implementation of the SFR-SIFT rejection classifier on a 2.3-GHz Pentium IV machine takes 4 seconds for matching a probe with a gallery of 466 faces and rejecting a subset of the gallery. Note that the SIFT generation code was in C++ [27].

5 Face Segmentation and Recognition

Fig. 10 shows the block diagram of our complete 3D face recognition algorithm including the rejection classifier and the final recognition process. During offline processing, the gallery is constructed from raw 3D faces. A single 3D face per individual is used. Each input 3D face is normalized, as described in Section 3. During online recognition, a probe from the test set is first preprocessed, as described in Section 3. Next, its SFR and SIFT features are computed and matched with those of the gallery to reject unlikely faces. The SFR-SIFT matching engine results in a vector of similarity scores s of size M (where M is the size of the gallery). The scores are normalized to a scale of 0 to 1 (0 being the best similarity) using (6)

$$\mathbf{s} = \frac{\mathbf{s} - min(\mathbf{s})}{max(\mathbf{s} - min(\mathbf{s})) - min(\mathbf{s} - min(\mathbf{s}))}. \qquad (6)$$

The only case when the denominator in (6) can be equal to zero is the highly unlikely case of the maximum similarity being equal to the minimum similarity. Gallery faces whose similarity is above a threshold are rejected. Selecting a threshold is a trade-off between accuracy and efficiency (or $\mathrm{Eff}(\psi)$). In our experiments, we used a threshold (equal to 0.29) so that the verification rate is 98 percent (neutral versus all case). The effectiveness $\mathrm{Eff}(\psi)$ of the rejection classifier was 0.036 at the 98 percent verification rate.

The remaining gallery faces are then verified using a more accurate matching engine. This matching engine is based on our *modified* ICP [5] algorithm (see Section 5.1 for details). To illustrate the sensitivity of the different regions of the 3D face to expressions, we register faces with a nonneutral expression to their respective 3D faces with a neutral expression and measure the variance in the depth of the corresponding pixels. Fig. 16 shows some sample 3D faces (first row), their variance due to facial expressions (second row), and a mask derived for each face (third row). The bright pixels in the second row correspond to greater facial expressions. In the third row, each mask represents a set of pixels of the face whose variance is less than the median variance of all the pixels. It is noticeable that, generally, the forehead, the region around the eyes, and the nose are the least affected by expressions (in 3D), whereas the cheeks and the mouth are the most affected. It is possible to derive a unique expression insensitive mask for each identity in the gallery and use it during the recognition process. This mask can be binary or it can assign a confidence value to each pixel depending upon its sensitivity to expressions. However, this approach is not possible using the FRGC v2.0 data as the data does not contain

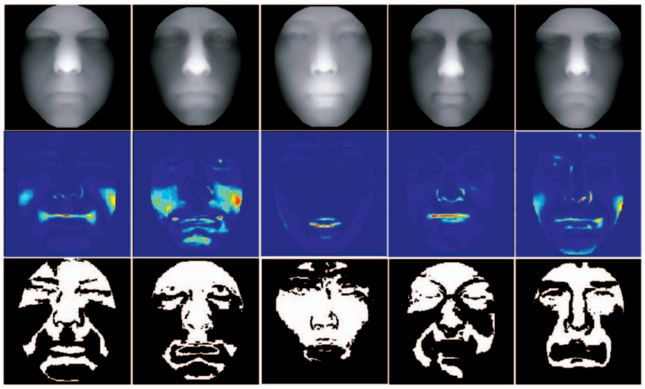


Fig. 16. Top: sample 3D faces. Center: variance of the 3D faces with expressions. Bottom: the expression insensitive binary mask of the 3D faces. Note that the nose and eyes-forehead regions are the least sensitive to expressions.

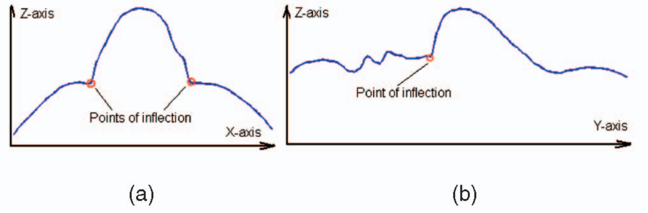


Fig. 17. Inflection points identified in (a) a horizontal slice and (b) a vertical slice of a 3D face.

faces with nonneutral expressions for every identity. Preliminary results using this approach are reported in our earlier work [34].

Given that the nose, the forehead, and the region around the eyes are the least sensitive to facial expressions in 3D faces (see Fig. 16), we segmented these features by automatically detecting the inflection points (see Fig. 17) around the nose tip. These inflection points are used to automatically segment the eyes-forehead and the nose regions from the face, as shown in Fig. 18. In our earlier work [32], [33], we segmented these regions from only the gallery faces and matched them to the complete probe face during recognition. This was mainly because the offline segmentation in [32], [33] was performed by manually selecting control points, and the online recognition process was meant to be fully automatic. However, in this paper, we have automated the segmentation process, and therefore, both the gallery and probe faces are segmented before matching.

The gallery faces were automatically segmented during offline processing. During online recognition, a part of the gallery is rejected using the SFR-SIFT rejection classifier. Next, the eyes-forehead and nose regions of the probe are segmented and individually matched with those of the remaining gallery faces using our modified ICP algorithm (see Section 5.1).

5.1 Matching

Matching is performed using a modified ICP algorithm [5]. ICP establishes correspondences between the *closest points* of two sets of 3D pointclouds and minimizes the distance error between them by applying a rigid transformation to one of the sets. This process is repeated *iteratively* until the distance error reaches a minimum saturation value. It also requires a prior coarse registration of the two pointclouds in order to avoid local minima. We use our automatic pose correction algorithm (Section 3.2) for this purpose. Our modified version of the ICP algorithm follows the same routine except that the correspondences are established along the z-axis only. The two pointclouds are mapped onto the xy-plane before correspondences are established between them. This way, points that are close in the xy-plane but far in the z-axis are still considered corresponding points. The distance error between such points provides useful information about the dissimilarity between two faces. However, points whose 2D distance in the xy-plane is more than the resolution of the faces ($1\ mm$) are not considered as corresponding points. Once the correspondences are established, the pointclouds are mapped back to their 3D coordinates, and the 3D distance error between them is minimized. This process is repeated until the error reaches a minimum saturation value.

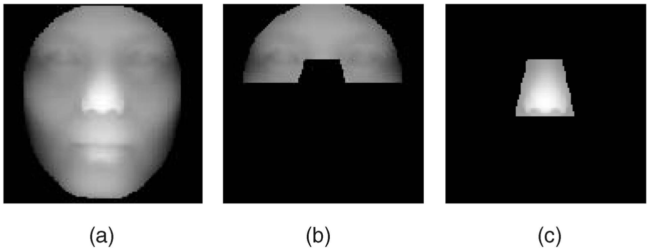


Fig. 18. A 3D face (a) is automatically segmented into the (b) eyes-forehead and (c) nose regions based on the inflection points around the nose.

Let $\mathbf{P} = [x_k, y_k, z_k]^\top$ (where $k = 1 \dots n_P$) and $\mathbf{G} = [x_k, y_k, z_k]^\top$ (where $k = 1 \dots n_G$) be the pointcloud of a probe and a gallery face, respectively. The projections of $\mathbf{P}$ and $\mathbf{G}$ on the xy-plane are given by $\widehat{\mathbf{P}} = [x_k, y_k]^\top$ and $\widehat{\mathbf{G}} = [x_k, y_k]^\top$, respectively. Let F be a function that finds the nearest point in $\widehat{\mathbf{P}}$ to every point in $\widehat{\mathbf{G}}$:

$$(\mathbf{c}, \mathbf{d}) = \mathrm{F}(\widehat{\mathbf{P}}, \widehat{\mathbf{G}}), \tag{7}$$

where $\mathbf{c}$ and $\mathbf{d}$ are vectors of size n_G each such that c_k and d_k contain, respectively, the index number and distance of the nearest point of $\widehat{\mathbf{P}}$ to the kth point of $\widehat{\mathbf{G}}$. For all k, find $g_k \in \mathbf{G}$ and $p_{c_k} \in \mathbf{P}$ such that $d_k < d_r$ (where d_r is the resolution of the 3D faces, equal to $1\ mm$ in our case). The resulting g_i correspond to p_i, for all $i = 1 \dots N$ (where N is the number of correspondences between $\mathbf{P}$ and $\mathbf{G}$). The distance error e to be minimized is given in (8). Note that e is the 3D distance error between the probe and the gallery as opposed to 2D distance. This error e is iteratively minimized and its final value is used as the similarity score between the probe and gallery face. To avoid local minima, a coarse to fine approach is used by initially setting a greater threshold for establishing correspondences and later bringing the threshold down to d_r. A higher initial threshold allows correspondences to be established between distant points in case the pose correction performed during normalization was not accurate

$$e = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{R} g_i + \mathbf{t} - p_i\|. \tag{8}$$

The rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ in (8) can be calculated using a number of approaches including Quaternions and the classic SVD (Singular Value Decomposition) method [1]. An advantage of the SVD method is that it can easily be generalized to any number of dimensions and is presented here for completeness. The mean of p_i and g_i is given by

$$\mu_p = \frac{1}{N}\sum_{i=1}^{N} p_i \quad \text{and} \tag{9}$$

$$\mu_g = \frac{1}{N}\sum_{i=1}^{N} g_i, \quad \text{respectively.} \tag{10}$$

The cross correlation matrix $\mathbf{K}$ between p_i and g_i is given by

$$\mathbf{K} = \frac{1}{N}\sum_{i=1}^{N}(g_i - \mu_g)(p_i - \mu_p)^\top. \tag{11}$$

Performing a Singular Value Decomposition of $\mathbf{K}$

$$\mathbf{UAV}^\top = \mathbf{K} \tag{12}$$

gives us two orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ and a diagonal matrix $\mathbf{A}$. The rotation matrix $\mathbf{R}$ can be calculated from the orthogonal matrices as

$$\mathbf{R} = \mathbf{VU}^\top, \tag{13}$$

whereas the translation vector $\mathbf{t}$ can be calculated as

$$\mathbf{t} = \mu_p - \mathbf{R}\mu_g. \tag{14}$$

$\mathbf{R}$ is a polar projection of $\mathbf{K}$. If $\det(\mathbf{R}) = -1$, this implies a reflection of the face in which case $\mathbf{R}$ is calculated using (15).

$$\mathbf{R} = \mathbf{V}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & det(\mathbf{UV}^\top) \end{bmatrix}\mathbf{U}^\top. \tag{15}$$

5.2 Fusion

Each matching engine results in a similarity matrix $\mathbf{S}_i$ (where i denotes a modality) of size $P \times M$ (where P is the number of tested probes, and M is the number of faces in the gallery). An element s_{rc} (at row r and column c) of a matrix $\mathbf{S}_i$ denotes the similarity score between probe number r and gallery face number c. Each row of an $\mathbf{S}_i$ represents an individual recognition test of probe number r. All the similarity matrices have a negative polarity in our case, that is, a smaller value of s_{rc} means high similarity. The individual similarity matrices are normalized before fusion. Since none of the similarity matrices had outliers, a simple min-max rule (16) was used for normalizing each row (recognition test) of a similarity matrix on a scale of 0 to 1

$$\mathbf{S}'_{ir} = \frac{\mathbf{S}_{ir} - \min(\mathbf{S}_{ir})}{\max(\mathbf{S}_{ir} - \min(\mathbf{S}_{ir})) - \min(\mathbf{S}_{ir} - \min(\mathbf{S}_{ir}))}, \tag{16}$$

$$\mathbf{S} = \prod_{i=1}^{n}\mathbf{S}'_i, \tag{17}$$

where $i = 1 \dots n$ (the number of modalities) and $r = 1 \dots P$ (the number of probes). Moreover, $\max(\mathbf{S}_{ir})$ and $\min(\mathbf{S}_{ir})$, respectively, represent the minimum and maximum value (that is, a scalar) of the entries of matrix $\mathbf{S}_i$ in row r. The normalized similarity matrices $\mathbf{S}'_i$ are then fused to get a combined similarity matrix $\mathbf{S}$. Two fusion techniques were tested, namely, multiplication and weighted sum. The multiplication rule (17) resulted in a slightly better verification rate but a significantly lower rank-one recognition rate. Therefore, we used the weighted sum rule (18) for fusion as it produced overall good verification and rank-one recognition results

$$\mathbf{S}_r = \sum_{i=1}^{n}\kappa_i\kappa_{ir}\mathbf{S}'_{ir}, \tag{18}$$

$$\kappa_{ir} = \frac{\text{mean}(\mathbf{S}'_{ir}) - \min(\mathbf{S}'_{ir})}{\text{mean}(\mathbf{S}'_{ir}) - \min_2(\mathbf{S}'_{ir})}. \tag{19}$$

In (18), κ_i is the confidence in modality i, and κ_{ir} is the confidence in recognition test r for modality i. In (19), $\min_2(\mathbf{S}'_{ir})$ is the second minimum value of $\mathbf{S}'_{ir}$. The final similarity matrix $\mathbf{S}$ is once again normalized using the min-max rule (20) resulting in $\mathbf{S}'$, which is used to calculate the combined performance of the used modalities

$$\mathbf{S}'_r = \frac{\mathbf{S}_r - \min(\mathbf{S}_r)}{\max(\mathbf{S}_r - \min(\mathbf{S}_r)) - \min(\mathbf{S}_r - min(\mathbf{S}_r))}. \tag{20}$$

When a rejection classifier is used, the resulting similarity matrices are sparse since a probe is matched with only a limited number of gallery faces. In this case, the gallery faces that are not tested are given a value of 1 in the normalized similarity matrix. Moreover, the confidence weight κ_{ir} is also set to 1 for every recognition trial. In some recognition trials, all faces are rejected but one. Since there is only one face left, it is declared as identified with a similarity of zero.

6 Results and Analysis

We present the results of three different variants of our algorithm. The first one is the multimodal hybrid face recognition (hereafter referred to as MMH_e) algorithm, as described in Fig. 10. The second variant comprises only the 3D eyes-forehead and nose matching engines and does not include the rejection classifier. This variant is referred to as the $R3D$ algorithm. The third variant fuses the 3D eyes-forehead and nose matching engines with the SFR-SIFT matching engine. This variant is referred to as the MMH_a algorithm. MMH_e is the most efficient, whereas MMH_a is the most accurate variant. The verification and identification results are presented in Sections 6.1 and 6.2 for comparison. Moreover, a comparison of MMH_a with existing algorithms is also provided in Section 6.3.

6.1 Verification Results

Fig. 19 shows the verification results of the MMH_e algorithm. Note that these results include the errors propagated from the rejection classifier. The verification rates at 0.001 FAR are 99.43 percent and 94.80 percent for probes with a neutral and a nonneutral expression, respectively. In the case of neutral versus neutral, the nose performs slightly better compared to the eyes-forehead because the latter is affected by hair. In the neutral versus nonneutral case, the results are the other way around, which is mainly because the eyes-forehead is comparatively less sensitive to expressions. Note that the nose region performed significantly better compared to our earlier experiments [33] because of our improved preprocessing techniques. These results also show that the human nose is an important and independent biometric, just like the human ear [49]. Its major advantages over the ear are that it

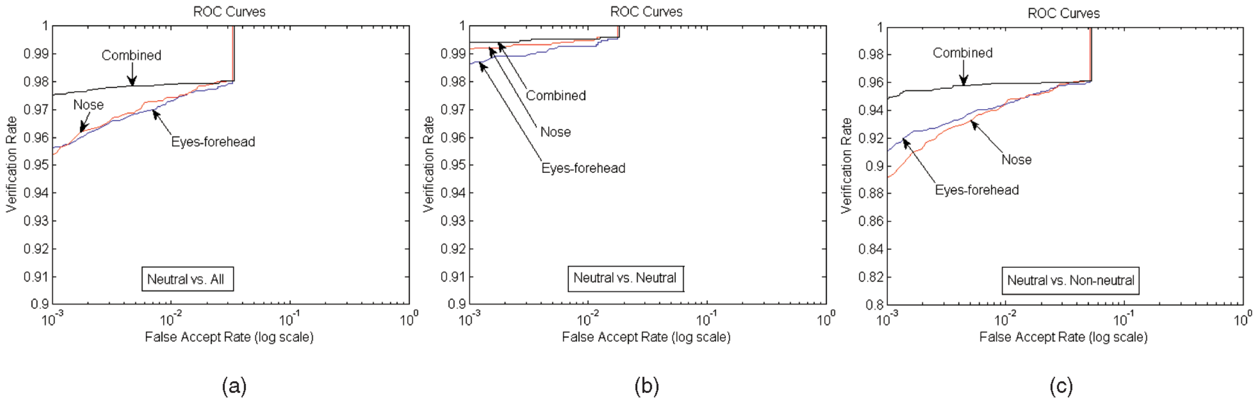


Fig. 19. Receiving operator characteristic (ROC) curves of the MMH_e algorithm, as shown in Fig. 10. The step in ROC curves is due to the use of a rejection classifier. The combined verification rate at 0.001 FAR for neutral versus all faces is 97.54 percent.

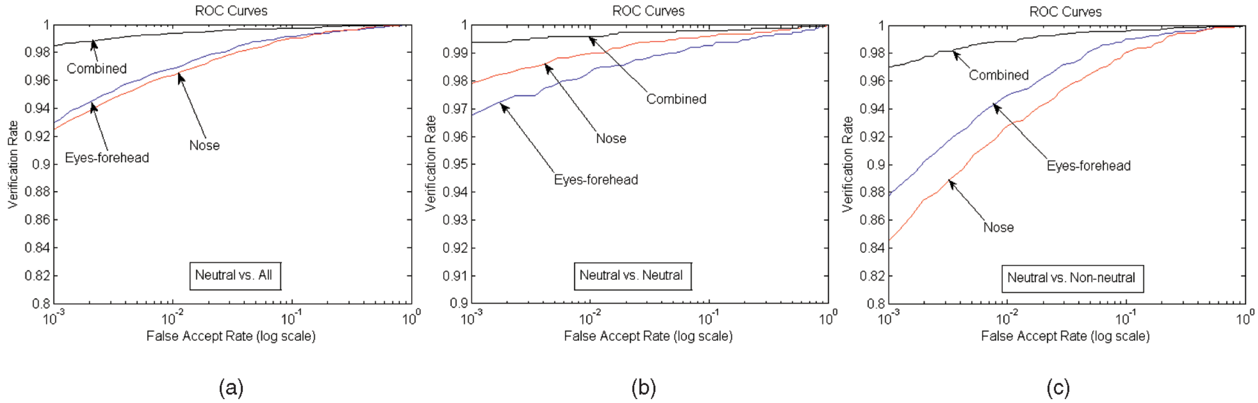


Fig. 20. ROC curves of the *R3D* algorithm. The combined verification rate at 0.001 FAR for neutral versus all faces is 98.5 percent.

can be observed from frontal, as well as profile, views and is unlikely to be occluded by hair.

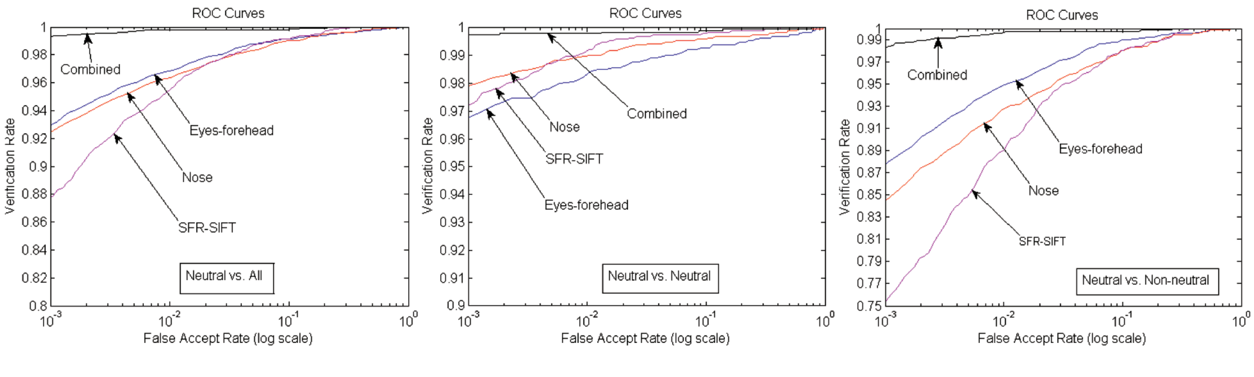
Fig. 20 shows the verification results of the *R3D* algorithm. In this case, the verification rates at 0.001 FAR are 99.38 percent and 97.00 percent for probes with a neutral and a nonneutral expression, respectively. Note that the verification results, with or without using a rejection classifier, for faces with neutral expressions are almost equal because at $\mathrm{Eff}(\psi) = 0.036$, the verification rate of the rejection classifier is very high (that is, 99.6 percent, see Fig. 15b). For faces with a nonneutral expression, the verification rate drops from 97.0 percent to 94.8 percent (by 2.2 percent) as a result of using the rejection classifier. There are many circumstances where a loss of 2 percent accuracy will be more than justified in order to achieve a nearly 30 fold improvement in runtime. An interesting point to note is that the individual performances of the eyes-forehead and the nose increase when the rejection classifier is used. For example, the verification rate of the eyes-forehead increases from 92.74 percent to 95.62 percent with rejection (Figs. 19a and 20a). This is because the rejection classifier increases the odds of recognition by reducing the effective gallery size. However, there is a probability that the classifier may also reject the correct identity, which is why the combined performance is deteriorated. The accuracy of the combined case is higher than the individual modalities, and therefore, the increase in odds of recognition is not of much use.

Fig. 21 shows the verification results of the MMH_a algorithm. The verification rate at 0.001 FAR in this case is 99.74 percent and 98.31 percent for faces with a neutral and a nonneutral expression, respectively.

Fig. 22 shows the performance of our *R3D* algorithm for the FRGC Experiment 3 (shape only), that is, when the query set (4,007 faces) is matched with the target set (4,007 faces). This amounts to 16 million comparisons each for the nose and eyes-forehead regions. The resulting two $(4,007 \times 4,007)$ similarity matrices are fused using a sum rule. As opposed to the remaining experiments in this paper, this is a one-to-one matching experiment and, therefore, the similarity matrices are not normalized in this case. The verification rates at 0.001 FAR of the eyes-forehead and the nose regions are 72.55 percent and 74.31 percent, respectively. The combined verification rate at 0.001 FAR is 86.6 percent, which is comparable to the best verification rate reported in [42] for Experiment 3 shape. The results of MMH_e and MMH_a for FRGC Experiment 3 are not reported due to the following reason. Recall that a rejection classifier quickly eliminates a large percentage of unlikely classes, and the accuracy required for rejection classifiers is less constraining compared to recognition classifiers [2]. The FRGC Experiment 3 is a one-to-one matching experiment, and hence, there is no need for a prior rejection phase.

6.2 Identification Results

Fig. 23 shows the results of the MMH_e algorithm. The identification rates in this case are 98.20 percent and 93.74 percent, respectively, for probes with a neutral and a nonneutral expression. These results are similar to the verification results in the sense that the nose performs slightly



(a) (b) (c)

Fig. 21. ROC curves of the MMH_a algorithm, that is, when the region-based matching engines are combined with the SFR-SIFT-based matching engine. No rejection is performed in this case. The combined verification rate at 0.001 FAR for neutral versus all faces is 99.32 percent.

better in the neutral versus neutral case and the eyes-forehead performs better in the neutral versus nonneutral case.

Fig. 24 shows the identification results of the $R3D$ algorithm. The identification rates in this case are 98.82 percent and 92.36 percent, respectively, for probes with a neutral and a nonneutral expression. The neutral versus neutral expression results are almost identical to Fig. 23b, whereas the identification rate drops from 93.74 percent to 92.36 percent (by 1.38 percent) as a result of using the rejection classifier. Fig. 25 shows the identification results of the MMH_a algorithm. The identification rates in this case are 99.02 percent and 95.37 percent for probes with neutral and nonneutral expressions, respectively.

6.3 Comparison with Other Algorithms

Table 1 compares the verification results of the MMH_a algorithm to others (including the baseline PCA performance), which used the FRGC v2.0 data. The FRGC benchmark is used for comparison, that is, the verification rate at 0.001 FAR. Note that our algorithm has the best 3D and multimodal performance and is the least sensitive to facial expressions. The results of Chang et al. [11] are not included in Table 1 since they did not report their verification rates. They reported a rank-one recognition rate of 91.9 percent for 3D faces (neutral versus all), which is lower than the 96.2 percent rank-one recognition rate of our algorithm (Fig. 24a).

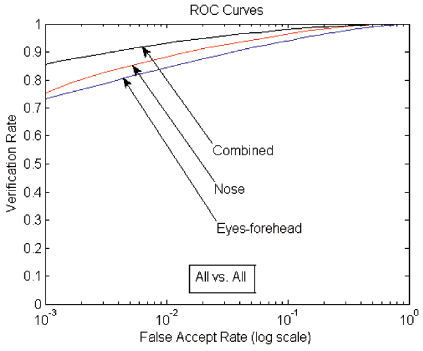


Fig. 22. ROC curves for all versus all case.

7 Limitations and Future Work

Our nose detection and pose correction algorithms assume that the input data contains a front view of a single face with small pose variations (±15 degrees) along the x-axis and the y-axis. However, pose variation along the z-axis can be between ±90 degrees. The accuracy of our nose detection algorithm is 98.3 percent (only 85 failures out of 4,950). The failures were mainly due to hair covering part of the face (see Fig. 26a) and, in a *few cases*, due to exaggerated expressions (for example, widely open mouth and inflated cheeks). Using a face detection algorithm (for example, [45]) in combination with a skin detection algorithm on the 2D colored images prior to nose detection can make the subsequent 3D nose detection simple and more accurate. The pose correction algorithm failed to correct the pose of 0.16 percent of the faces (only eight out of 4,950) along the z-axis. Hair was also a source of error during pose correction, as shown in Fig. 26b. Hair also caused problems in calculating the SFR and during the final verification process. Most of the false positives during the eyes-forehead and nose matching occurred due to hair and exaggerated expressions. Fig. 27 shows six example probes and their corresponding gallery faces, which could not be recognized by our algorithm. Examples of challenging faces correctly recognized by our algorithm can be seen in Fig. 7 (except for the second last face in the third row). A skin detection algorithm could be useful to overcome the limitations due to hair [33]. However, applying it before pose correction will result in missing regions from the face (because they were covered by hair) leading to an incorrect pose. Another source of error was the inherent coregistration error between the 2D and 3D faces in the FRGC v2.0 data, which resulted in an incorrect region being cropped from the 2D face (see Fig. 26c).

In our future work, we intend to use skin detection in our algorithm and fill in the missing regions by using morphable models and facial symmetry. We also intend to use a more robust illumination normalization algorithm to improve the recognition performance on 2D images. Finally, we aim to extend our algorithms to be able to automatically detect profile views and perform fully automatic face recognition on them.

8 Conclusion

We presented a fully automatic multimodal hybrid face recognition algorithm and demonstrated its performance on

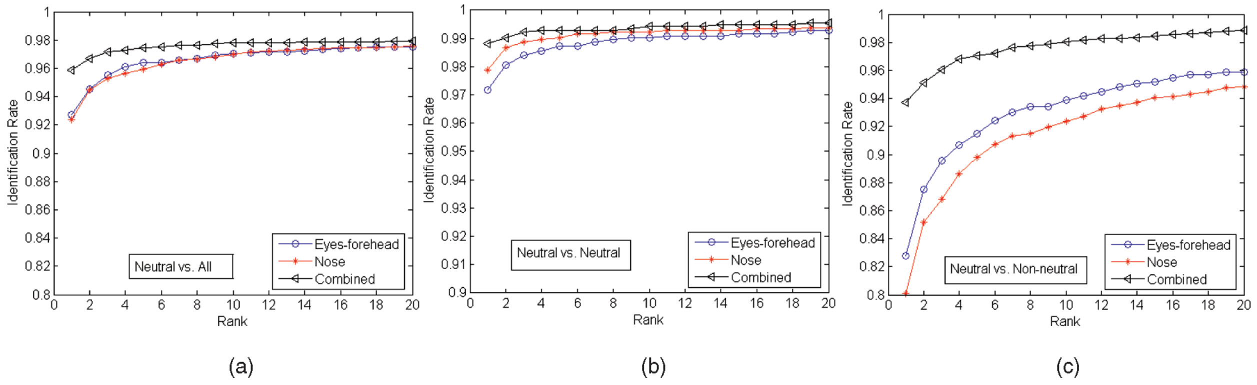


Fig. 23. Identification results of the MMH_e algorithm, as shown in Fig. 10. The combined rank-one identification rate for neutral versus all faces is 95.91 percent.

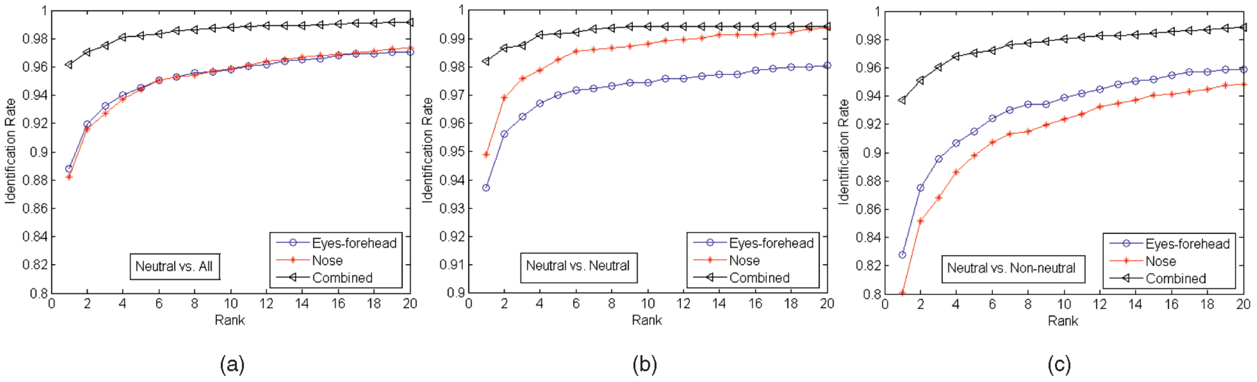


Fig. 24. Identification results of the *R3D* algorithm. The combined rank-one identification rate for neutral versus all faces is 96.2 percent.

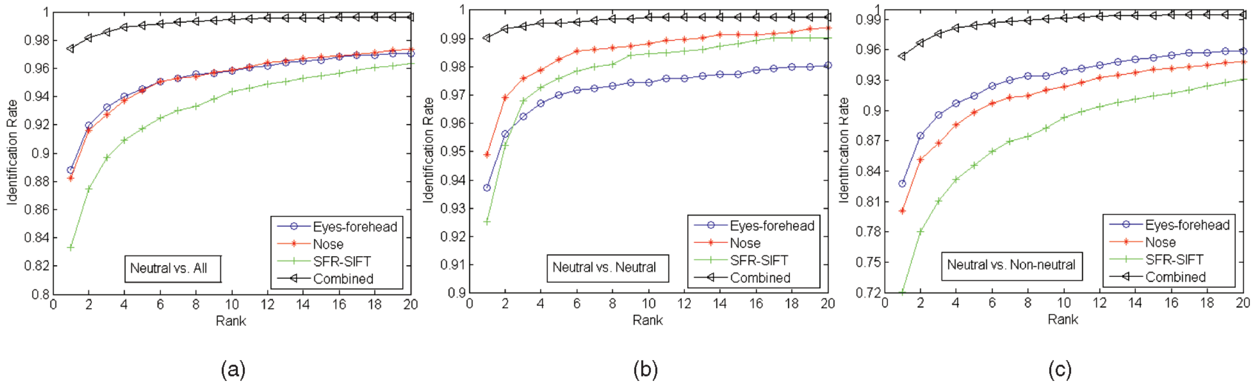


Fig. 25. Identification results of the MMH_a algorithm, that is, when the region-based matching engines are combined with the SFR-SIFT-based matching engine. No rejection is performed in this case. The combined rank-one identification rate for neutral versus all faces is 97.37 percent.

the FRGC v2.0 data set. Several novelties were presented while addressing major problems in the area of 3D and multimodal face recognition. Our contributions include

1. a fully automatic pose correction algorithm,
2. an SFR for 3D faces,
3. a novel SFR-SIFT-based rejection classifier, and
4. a region-based matching algorithm for expression robust face recognition.

Although these algorithms have been applied to 3D face recognition, they can easily be generalized to other 3D shapes. In addition to these novelties, we, for the first time in the literature, successfully used the 3D nose as an independent biometric. Furthermore, we also presented the first ever comprehensive study (that is, using a large database) on the use of SIFT descriptors for 2D face recognition under illumination and expression variations. We addressed three major challenges, namely, automation, efficiency, and robustness to facial expressions. The performance of our algorithms was demonstrated on the largest publicly available corpus of 3D face data. The performance of the SFR-SIFT rejection classifier was 0.036, which amounts to 27.78 times improvement in recognition time. Our multimodal hybrid recognition algorithms achieve 99.74 percent and 98.31 percent verification rates at 0.001 FAR for faces with a neutral and a nonneutral expression, respectively. The identification rates for the same were 99.02 percent and 95.37 percent. In terms of accuracy, these results are slightly better than any

TABLE 1
Verification Rates at 0.001 FAR Using the FRGC v2.0 Data

	Neutral vs. All		Neutral vs. Neutral		Neutral vs. Non−neutral	
	3D	Multimodal	3D	Multimodal	3D	Multimodal
This paper	98.5 %	99.3 %	99.4 %	99.7 %	97 %	98.3 %
Maurer et al. [29]	86.5 %	95.8 %	97.8%	99.2 %	NA	NA
Husken et al. [18]	89.5 %	97.3 %	NA	NA	NA	NA
Passalis et al. [38]	85.1 %	NA	94.9 %	NA	79.4 %	NA
FRGC baseline	45 %	54 %	NA	82 %	40 %	43 %

Our algorithm has the best 3D and multimodal performance and is the least sensitive to facial expressions (NA: Not Available).

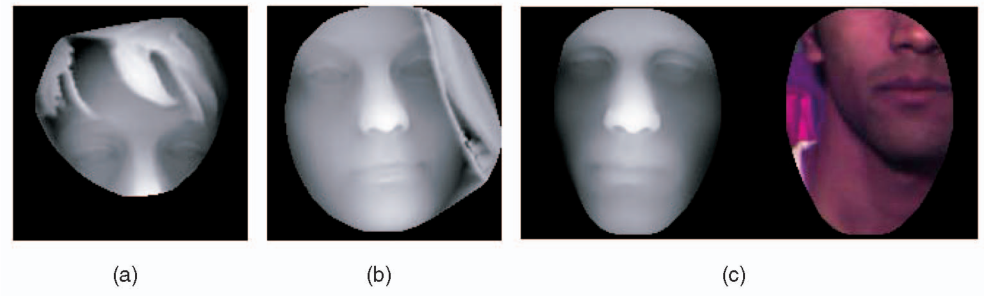


Fig. 26. An example of (a) incorrect nose detection, (b) incorrect pose correction due to hair, and (c) registration error in the FRGC data.

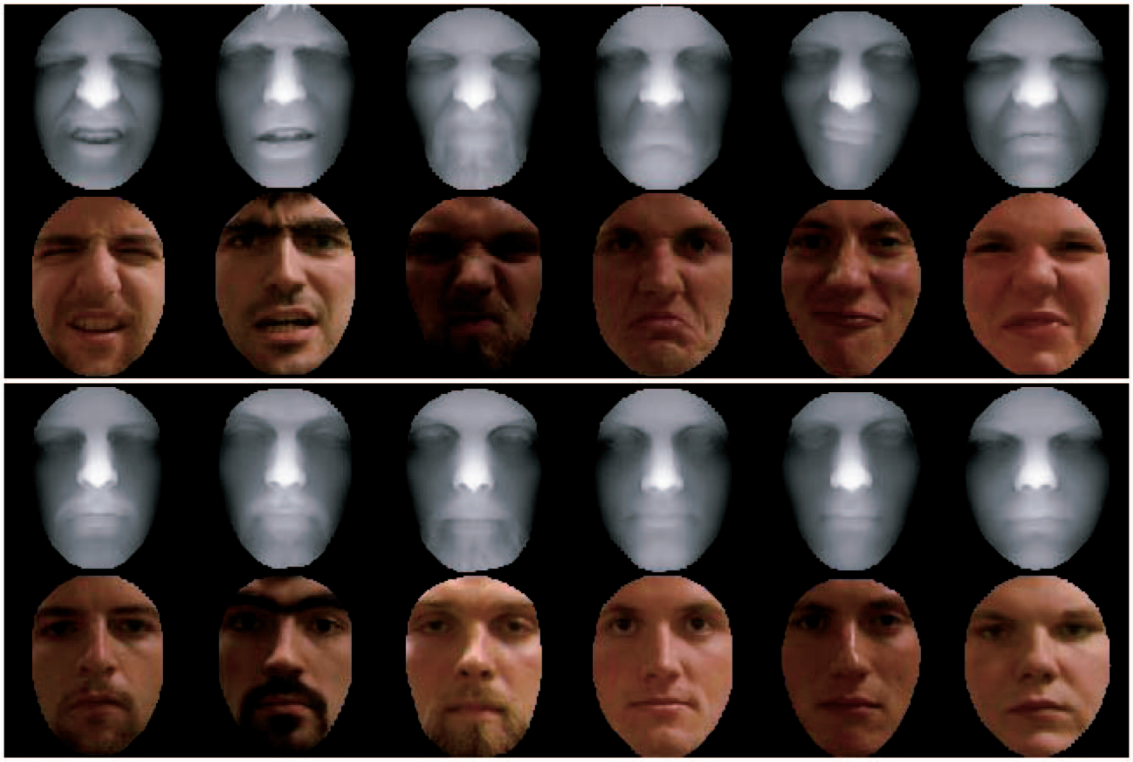


Fig. 27. Example probes that could not be recognized correctly (top) and their corresponding gallery faces (bottom).

previously published results. This is quite understandable as there was little room for improvement. The individual verification rate at 0.001 FAR of our 3D region-based matching algorithm alone is 98.5 percent (Fig. 20a), which is a strong indicator of the potential of 3D face recognition. Our results show that the eyes-forehead and nose regions of a face contain the maximum discriminating features important for the expression of robust face recognition.

ACKNOWLEDGMENTS

The authors would like to thank the organizers of FRGC, Jonathon Phillips, Kevin Bowyer, and Patrick Flynn, for providing the face data. We would also like to thank David Lowe for making the SIFT code available. This work was sponsored by the Australian Research Council (ARC) Discovery Grant DP0664228.

REFERENCES

[1] K. Arun, T. Huang, and S. Blostein, "Least-Squares Fitting of Two 3-D Point Sets," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 9, no. 5, pp. 698-700, 1987.

[2] S. Baker and S.K. Nayar, "Pattern Rejection," *Proc. IEEE Computer Vision and Pattern Recognition,* pp. 544-549, 1996.

[3] M.S. Bartlett, H.M. Lades, and T. Sejnowski, "Independent Component Representation for Face Recognition," *Proc. SPIE Symp. Electronic Imaging,* pp. 528-539, 1998.

[4] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, pp. 711-720, 1997.

[5] P.J. Besl and N.D. McKay, "Reconstruction of Real-World Objects via Simultaneous Registration and Robust Combination of Multiple Range Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 14, no. 2, pp. 239-256, Feb. 1992.

[6] Biometric Consortium, http://www.biometrics.org, 2004.

[7] K.W. Bowyer, K. Chang, and P. Flynn, "A Survey of Approaches and Challenges in 3D and Multi-Modal 3D + 2D Face Recognition," *Computer Vision and Image Understanding,* vol. 101, pp. 1-15, 2006.

[8] A.M. Bronstein, M.M. Bronstein, and R. Kimmel, "Three-Dimensional Face Recognition," *Int'l J. Computer Vision,* vol. 64, no. 1, pp. 5-30, 2005.

[9] K.I. Chang, K.W. Bowyer, and P.J. Flynn, "Face Recognition Using 2D and 3D Facial Data," *Proc. Workshop Multimodal User Authentication,* pp. 25-32, 2003.

[10] K.I. Chang, K.W. Bowyer, and P.J. Flynn, "Multi-Modal 2D and 3D Biometrics for Face Recognition," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures,* pp. 187-194, 2003.

[11] K.I. Chang, K.W. Bowyer, and P.J. Flynn, "ARMS: Adaptive Rigid Multi-Region Selection for Handling Expression Variation in 3D Face Recognition," *Proc. IEEE Workshop Face Recognition Grand Challenge (FRGC) Experiments,* 2005.

[12] C.S. Chua and R. Jarvis, "Point Signatures: A New Representation for 3D Object Recognition," *Int'l J. Computer Vision,* vol. 25, no. 1, pp. 63-85, 1997.

[13] C. Chua, F. Han, and Y. Ho, "3D Human Face Recognition Using Point Signatures," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures,* pp. 233-238, 2000.

[14] I.J. Cox, J. Ghosn, and P.N. Yianilos, "Feature-Based Face Recognition Using Mixture Distance," *Proc. IEEE Computer Vision and Pattern Recognition,* pp. 209-216, 1996.

[15] Geometrix, http://www.geometrix.com/, 2006.

[16] R.C. Gonzalez and R.E. Woods, *Digital Image Processing.* Addison-Wesley, 1992.

[17] J. Huang, B. Heisele, and V. Blanz, "Component-Based Face Recognition with 3D Morphable Models," *Proc. Int'l Conf. Audio- and Video-Based Person Authentication,* 2003.

[18] M. Huskën, M. Brauckmann, S. Gehlen, and C. Malsburg, "Strategies and Benefits of Fusion of 2D and 3D Face Recognition," *Proc. IEEE Workshop Face Recognition Grand Challenge Experiments,* 2005.

[19] A.K. Jain, L. Hong, and S. Pankanti, "Biometric Identification," *Comm. ACM,* vol. 43, no. 2, pp. 91-98, 2000.

[20] A.K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 14, no. 1, pp. 4-20, 2004.

[21] A.E. Johnson and M. Hebert, "Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 5, pp. 674-686, May 1999.

[22] P. Kovesi, "MATLAB and Octave Functions for Computer Vision and Image Processing," http://people.csse.uwa.edu.au/pk/Research/MatlabFns/index.html, 2006.

[23] A. Lanitis, C. Taylor, and T. Cootes, "Automatic Face Identification System Using Flexible Appearance Models," *Image and Vision Computing,* vol. 13, pp. 393-401, 1995.

[24] D. Lin and X. Tang, "Recognize High Resolution Faces: From Macrocosm to Microcosm," *Proc. IEEE Computer Vision and Pattern Recognition,* pp. 1355-1362, 2006.

[25] C. Liu and H. Wechsler, "Evolutionary Pursuit and Its Application to Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, pp. 570-582, 2000.

[26] D. Lowe, "Distinctive Image Features from Scale-Invariant Key Points," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[27] D. Lowe, "Demo Software: SIFT Keypoint Detector," http://www.cs.ubc.edu.ca/~lowe/, 2006.

[28] X. Lu, A.K. Jain, and D. Colbry, "Matching 2.5D Scans to 3D Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 1, pp. 31-43, Jan. 2006.

[29] T. Maurer, D. Guigonis, I. Maslov, B. Pesenti, A. Tsaregorodtsev, D. West, and G. Medioni, "Performance of Geometrix ActiveID 3D Face Recognition Engine on the FRGC Data," *Proc. IEEE Workshop Face Recognition Grand Challenge Experiments,* 2005.

[30] G. Medioni and R. Waupotitsch, "Face Recognition and Modeling in 3D," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures,* pp. 232-233, 2003.

[31] A.S. Mian, M. Bennamoun, and R.A. Owens, "A Novel Representation and Feature Matching Algorithm for Automatic Pairwise Registration of Range Images," *Int'l J. Computer Vision,* vol. 66, pp. 19-40, 2006.

[32] A.S. Mian, M. Bennamoun, and R.A. Owens, "Region-Based Matching for Robust 3D Face Recognition," *Proc. British Machine Vision Conf.,* vol. 1, pp. 199-208, 2005.

[33] A.S. Mian, M. Bennamoun, and R.A. Owens, "2D and 3D Multimodal Hybrid Face Recognition," *Proc. European Conf. Computer Vision,* part III, pp. 344-355, 2006.

[34] A.S. Mian, M. Bennamoun, and R.A. Owens, "Automatic 3D Face Detection, Normalization and Recognition," *Proc. Third Int'l Symp. 3D Data Processing, Visualization and Transmission,* 2006.

[35] A.S. Mian, M. Bennamoun, and R.A. Owens, "Three-Dimensional Model-Based Object Recognition and Segmentation in Cluttered Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 10, pp. 1584-1601, Oct. 2006.

[36] Minolta 3D Digitizers, *Non-Contact 3D Laser Scanner,* http://www.minolta3d.com, 2006.

[37] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, pp. 696-710, 1997.

[38] G. Passalis, I. Kakadiaris, T. Theoharis, G. Tederici, and N. Murtaza, "Evaluation of 3D Face Recognition in the Presence of Facial Expressions: An Annotated Deformable Model Approach," *Proc. IEEE Workshop Face Recognition Grand Challenge Experiments,* 2005.

[39] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. IEEE Computer Vision and Pattern Recognition,* 1994.

[40] P.J. Phillips, "Support Vector Machines Applied to Face Recognition," *Proc. 1998 Conf. Advances in Neural Information Processing Systems,* vol. 11, pp. 803-809, 1998.

[41] P.J. Phillips, P.J. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proc. IEEE Computer Vision and Pattern Recognition,* 2005.

[42] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, and W. Worek, "Preliminary Face Recognition Grand Challenge Results," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* 2006.

[43] F. Samaria and S. Young, "HMM Based Architecture for Face Identification," *Int'l J. Image and Vision Computing,* vol. 12, pp. 537-583, 1994.

[44] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, 1991.

[45] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision,* vol. 57, no. 2, pp. 137-154, 2004.

[46] Y. Wang, C. Chua, and Y. Ho, "Face Recognition From 2D and 3D Images," *Proc. Int'l Conf. Audio and Video-Based Person Authentication,* pp. 26-31, 2001.

[47] L. Wiskott, J. Fellous, and C. Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, pp. 775-779, 1997.

[48] C. Xu, Y. Wang, T. Tan, and L. Quan, "Automatic 3D Face Recognition Combining Global Geometric Features with Local Shape Variation Information," *Proc. IEEE Int'l Conf. Pattern Recognition,* pp. 308-313, 2004.

[49] P. Yan and K. Bowyer, "Empirical Evaluation of Advanced Ear Biometrics," *Proc. IEEE Computer Vision and Pattern Recognition Workshops,* pp. 308-313, 2004.

[50] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys,* pp. 399-458, 2003.

Ajmal S. Mian received the BE degree in avionics from the College of Aeronautical Engineering, Nadirshaw Edulji Dinshaw (NED) University, Pakistan, in 1993, the MS degree in information security from the National University of Sciences and Technology, Pakistan, in 2003, and the PhD degree in computer science with distinction from the University of Western Australia in 2006. He is currently a research fellow in the School of Computer Science and Software Engineering, University of Western Australia. His research interests include computer vision, pattern recognition, multimodal biometrics, and information security. He worked on a number of engineering and R&D projects related to radar data processing and communication jamming and antijamming techniques before he was nominated for a master's degree.

Mohammed Bennamoun received the MSc degree in control theory from Queen's University, Kingston, Canada, and the PhD degree in computer vision from Queen's/QUT in Brisbane, Australia. He lectured robotics at Queen's and then joined QUT in 1993 as an associate lecturer. He then became a lecturer in 1996 and a senior lecturer in 1998 at QUT. In January 2003, he joined the School of Computer Science and Software Engineering, University of Western Australia as an associate professor. He was also the director of a research center from 1998 to 2002. His research interests include control theory, robotics, obstacle avoidance, object recognition, artificial neural networks, signal/image processing, and computer vision. He is the coauthor of the book *Object Recognition: Fundamentals and Case Studies* (Springer-Verlag, 2001). He published more than 100 journal and conference publications. He served as a guest editor for a couple of special issues in International journals such as the *International Journal of Pattern Recognition and Artificial Intelligence* (IJPRAI). He was selected to give conference tutorials at the European Conference on Computer Vision (ECCV '02) and the International Conference on Acoustics Speech and Signal Processing (ICASSP) in 2003. He organized several special sessions for conferences; the latest was for the IEEE International Conference in Image Processing (ICIP) held in Singapore in 2004. He also contributed in the organization of many local and international conferences.

Robyn Owens received the BSc(Hons) degree in mathematics at University of Western Australia (UWA) in 1974 and the MSc (1976) and the PhD (1980) degrees in mathematics at Oxford University. She spent three years in Paris at l'Universite de Paris-Sud, Orsay, continuing research in mathematical analysis before returning to UWA in 1982 to work as a research mathematician on the Automated Sheep Shearing Project. Since then, she has lectured in the Mathematics Department and, in 1986, joined the Department of Computer Science and Software Engineering after a six month visiting lectureship at Berkeley. She is currently the dean of graduate studies at UWA. Her research interests include feature detection in visual images and shape measurement and representation.