Int J Comput Vis (2008) 79: 1–12
DOI 10.1007/s11263-007-0085-5

Keypoint Detection and Local Feature Matching for Textured 3D Face Recognition

Ajmal S. Mian · Mohammed Bennamoun · Robyn Owens

Received: 6 March 2007 / Accepted: 20 August 2007 / Published online: 25 September 2007

Abstract Holistic face recognition algorithms are sensitive to expressions, illumination, pose, occlusions and makeup. On the other hand, feature-based algorithms are robust to such variations. In this paper, we present a feature-based algorithm for the recognition of textured 3D faces. A novel keypoint detection technique is proposed which can repeatably identify keypoints at locations where shape variation is high in 3D faces. Moreover, a unique 3D coordinate basis can be defined locally at each keypoint facilitating the extraction of highly descriptive pose invariant features. A 3D feature is extracted by fitting a surface to the neighborhood of a keypoint and sampling it on a uniform grid. Features from a probe and gallery face are projected to the PCA subspace and matched. The set of matching features are used to construct two graphs. The similarity between two faces is measured as the similarity between their graphs. In the 2D domain, we employed the SIFT features and performed fusion of the 2D and 3D features at the feature and score-level. The proposed algorithm achieved 96.1% identification rate and 98.6% verification rate on the complete FRGC v2 data set.

Keywords Feature-based face recognition · Keypoint detection · Invariant features · Feature-level fusion

This work is supported by ARC grant number DP0664228.

A.S. Mian (✉) · M. Bennamoun · R. Owens
School of Computer Science and Software Engineering,
The University of Western Australia, 35 Stirling Highway,
Crawley, WA 6009, Australia
e-mail: ajmal@csse.uwa.edu.au

M. Bennamoun
e-mail: bennamou@csse.uwa.edu.au

R. Owens
e-mail: robyn.owens@uwa.edu.au

1 Introduction

The human face has emerged as one of the most promising biometrics due to the social acceptability and non-intrusiveness of its measurement through imaging. It requires minimal or no cooperation from the subject making it ideal for surveillance and applications where customer satisfaction is important. However, machine recognition of faces is very challenging because the distinctiveness of facial biometrics is quite low compared to other biometrics (e.g. fingerprints and iris) (Jain et al. 2004). Moreover, changes caused by expressions, illumination, pose, occlusions and facial makeup (e.g. beard) impose further challenges on accurate face recognition.

A comprehensive survey of basic face recognition algorithms is given by Zhao et al. (2003) who categorize face recognition into holistic, feature-based and hybrid matching algorithms. Holistic matching algorithms basically extract global features from the entire face. Eigenfaces (Turk and Pentland 1991) and Fisherfaces (Belhumeur et al. 1997) are well known examples of holistic face recognition algorithms. Feature-based matching algorithms extract local features or regions such as the eyes and nose and then match these features or their local statistics for recognition. One example of this category is the region-based 3D matching algorithm (Mian et al. 2007) which matches the 3D point-clouds of the eyes-forehead and the nose regions separately and fuse the results at the score-level. Another example is face recognition using local boosted features (Jones and Viola 2003) which matches rectangular regions from facial images at different locations, scales and orientations. Hybrid matching methods use a combination of global and local-features for face recognition e.g. Huang et al. 2003).

One limitation of holistic matching is that it requires accurate normalization of the faces according to pose, illu-

mination and scale. Variations in these factors can affect the global features extracted from the faces leading to inaccuracies in the final recognition. Normalization is usually performed by manually identifying landmarks on the faces which makes the whole process semi-automatic. Manual normalization is also imperfect as it is commonly performed on the basis of only a few landmarks (3 to 5) and humans cannot identify landmarks with subpixel accuracy. Replacing this manual process with an automatic landmark identification algorithm usually deteriorates the final recognition results. Moreover, global features are also sensitive to facial expressions and occlusions. Feature-based matching algorithms have an advantage over holistic matching algorithms because they are robust to variations in pose, illumination, scale, expressions and occlusions.

Multimodal 2D-3D face recognition provides more accurate results than either of the individual modalities alone (Bowyer et al. 2006). An up to date survey of 3D and multimodal face recognition is given by Bowyer et al. (2006) who argue that 3D face recognition has the potential to overcome the limitations of its 2D counterpart. 3D data can assist in the normalization as well as recognition of faces. For example, 3D face data can be used for illumination normalization (Al-Osaimi et al. 2006) and pose correction (Mian et al. 2006c) of 2D faces. In the recognition phase, additional features can be extracted from the 3D faces and fused with the 2D features at the score or feature-level for greater accuracy. Bowyer et al. (2006) stress upon the need for better 3D face recognition algorithms which are more robust to variations. However, they also agree that multiple algorithms or classifiers can increase performance. For example, Gokberk et al. (2005) reported an increase in recognition performance by combining the results of multiple 3D face recognition algorithms.

Many 3D face recognition approaches are based on the Iterative Closest Point (ICP) algorithm (Besl and McKay 1992) or its modified versions (Mian et al. 2007). There are two major advantages of ICP based approaches. Firstly, perfect normalization of the faces is not required as the algorithm iteratively corrects registration errors while matching. Secondly, a partial region of a face can be matched with a complete face. The latter advantage has been exploited to avoid facial expressions (Mian et al. 2006a, 2007) and to handle pose variations by matching 2.5D scans to complete face models (Lu et al. 2006). On the downside, the major disadvantage of ICP is that it is an iterative algorithm and is therefore computationally very expensive. Moreover, ICP does not extract any feature from the face which rules out the possibilities of feature-level fusion and indexing to speed up the matching process. Unless another classifier and/or modality is used to perform indexing or prior rejection of unlikely faces from the gallery (Mian et al. 2006c), ICP based algorithms must perform a brute force matching thereby making the recognition time linear to the gallery size. Matching expression insensitive regions of the face is a potentially useful approach to overcome the sensitivity of ICP to expressions. However, deciding upon such regions is a problem worth exploring as such regions may not only vary between different persons but between different expressions as well.

Our literature review leads us to the following conclusions: (1) Local features are more robust compared to global features; (2) Multimodal 2D-3D face recognition is more accurate than either of the individual modalities; (3) The face recognition literature can always benefit from not only better but more number of face recognition algorithms because their combination is likely to improve performance; (4) Unlike ICP, it is more advantageous to extract features from the 3D face in order to facilitate feature-level fusion and indexing.

Based on the above motivations, we propose an algorithm for textured 3D face recognition using local features extracted separately from the 3D shape and texture of the face. In the 3D domain, a novel approach for keypoint detection is proposed. The identification of keypoints is repeatable and allows for the extraction of highly descriptive 3D features at the keypoints. Features are extracted with reference to local object-centered 3D coordinates which are uniquely defined at each keypoint using its neighboring surface. The repeatability of the keypoint locations and reference frames provides pose invariance to the 3D features. Each feature is extracted by fitting a surface to the neighborhood of a keypoint and sampling it on a uniform grid. In order to achieve robustness to noise, approximation is used for surface fitting rather than interpolation. Multiple features are extracted from each gallery face and projected to the PCA subspace. The dimensionality of the feature vectors is reduced by considering only the most significant components. During recognition, features are extracted at keypoints on the probe. These features are also projected to the PCA subspace and matched with those in the gallery. The set of matching features from a probe and gallery face are individually meshed to form graphs. A spatial constraint is used to remove false matches and the remaining ones are used to calculate the similarity measure between the two faces. The similarity measure is not only based on the average error between the corresponding features but also takes into account the error between the two graphs and the total number of correct matches between the faces.

We also employed the SIFT features in the 2D domain and fused the two modalities at the score and feature-level. In the former case, the two feature sets (2D and 3D) are extracted and matched independently and their similarity scores are fused. In the latter case, both the 2D and 3D features are extracted at the same keypoints on the face and projected to their respective PCA subspaces. The dimensionality of each feature is reduced and each feature vector

is normalized so that its variance along every dimension is equal. Finally, the feature vectors are normalized to a unit magnitude and concatenated to form a multimodal (2D-3D) feature vector. The same algorithm as described above is also used for matching the multimodal features. The proposed algorithm was tested on the FRGC v2 (Phillips et al. 2005) data set and achieved identification rates of 99.38% and 92.11% for probes with neutral and non-neutral expressions, respectively. The verification rates at 0.001 FAR for the same were 99.85% and 96.62%.

Preliminary results of our algorithm have been published in (Mian et al. 2006d). However, a number of extensions have been made since then which resulted in a significant improvement in the accuracy (from 89.5% to 99.38% 3D face identification rate) and efficiency of the algorithm. These extensions include a keypoint detection algorithm, 3D coordinate basis derivation from single keypoints, projection of the features to PCA subspace, feature-level fusion, using a more sophisticated graph matching approach and performing experiments on the complete FRGC v2 data set.

The rest of this paper is organized as follows. Section 2 gives details of the keypoint identification algorithm. Section 3 explains the extraction of local 3D features. Section 4.1 gives a brief overview of the SIFT features. Feature-level fusion is described in Sect. 4.2. Details of the matching algorithm are given in Sect. 5. Results are reported in Sect. 6. Finally, discussion and conclusions are given in Sects. 7 and 8, respectively.

2 Keypoint Detection

The aim of keypoint detection is to determine points on a surface (a 3D face in our case) which can be identified with high repeatability in different range images of the same surface in the presence of noise and pose variations. In addition to repeatability, the features extracted from these keypoints should be sufficiently distinctive in order to facilitate accurate matching. The keypoint identification technique proposed in this paper is simple yet robust due to its repeatability and the descriptiveness of the features extracted at these keypoints.

The input to our algorithm is a point cloud of a face $\mathbf{F} = [x_i \ y_i \ z_i]^T$, where $i = 1, \ldots, n$. The input face is sampled at uniform (x, y) intervals (4 mm in our case) and at each sample point p, a local surface is cropped from the face using a sphere of radius r_1 centered at p. The value of r_1 decides the degree of locality of the extracted feature and offers a trade off between descriptiveness and sensitivity to variations e.g. due to expressions. The smaller the value of r_1, the less the sensitivity of the local feature to variations and vice versa. However, on the downside a small value of r_1 will also decrease the descriptiveness of the feature.

Let $\mathbf{L}_j = [x_j \ y_j \ z_j]^T$ (where $j = 1, \ldots, n_l$) be the points in the region cropped by the sphere of radius r_1 centered at p. The mean vector $\mathbf{m}$ and the covariance matrix $\mathbf{C}$ of $\mathbf{L}$ are given by

$$\mathbf{m} = \frac{1}{n_l} \sum_{j=1}^{n_l} \mathbf{L}_j, \quad \text{and} \tag{1}$$

$$\mathbf{C} = \frac{1}{n_l} \sum_{j=1}^{n_l} \mathbf{L}_j \mathbf{L}_j^T - \mathbf{m}\mathbf{m}^T, \tag{2}$$

where $\mathbf{L}_j$ is the jth column of $\mathbf{L}$. Performing Principal Component Analysis on the covariance matrix $\mathbf{C}$ gives the matrix $\mathbf{V}$ of eigenvectors

$$\mathbf{C}\mathbf{V} = \mathbf{D}\mathbf{V}, \tag{3}$$

where $\mathbf{D}$ is the diagonal matrix of the eigenvalues of $\mathbf{C}$. The matrix $\mathbf{L}$ can be aligned with its principal axes using (4), known as the Hotelling transform (Gonzalez and Woods 1992):

$$\mathbf{L}'_j = \mathbf{V}(\mathbf{L}_j - \mathbf{m}) \quad \{j = 1, \ldots, n_l\}. \tag{4}$$

Let $\mathbf{L}'_x$ and $\mathbf{L}'_y$ represent the x and y components of the point cloud $\mathbf{L}'$ i.e.

$$\mathbf{L}'_x = x_j \quad \text{and} \quad \mathbf{L}'_y = y_j \quad \text{where } \{j = 1, \ldots, n_l\}, \tag{5}$$

$$\delta = \max(\mathbf{L}'_x) - \min(\mathbf{L}'_x) - (\max(\mathbf{L}'_y) - \min(\mathbf{L}'_y)). \tag{6}$$

In (6), δ is the difference between the first two principal axes of the local region. The value of δ will be zero if $\mathbf{L}'$ is planar or spherical. However, if there is unsymmetrical variation in the depth of $\mathbf{L}'$, then δ will have a non-zero value proportional to the variation. This depth variation is an indication that $\mathbf{L}'$ contains descriptive information.

The value of δ is always non-negative and if it is greater than a threshold (i.e. $\delta \geq t_1$), p is taken as a keypoint, otherwise it is rejected. The threshold t_1 governs the total number of keypoints. Taking $t_1 = 0$ will result in every point ending up as a keypoint and as the value of t_1 increases the total number of keypoints will decrease. There are only two thresholds in the keypoint detection namely r_1 and t_1 which are empirically chosen as $r_1 = 20$ mm and $t_1 = 2$ mm. However, the algorithm is not sensitive to these parameters and small changes in these values will not have a significant effect on its performance. The values of r_1 and t_1 depend on the scale of human faces. This keypoint detection technique is generic and can be extended to other 3D shapes and objects. However, the values of r_1 and t_1 will change accordingly relative to the scale of the objects to be recognized.

Figure 1 shows some examples of keypoints identified on three different range images each of four individuals. We can

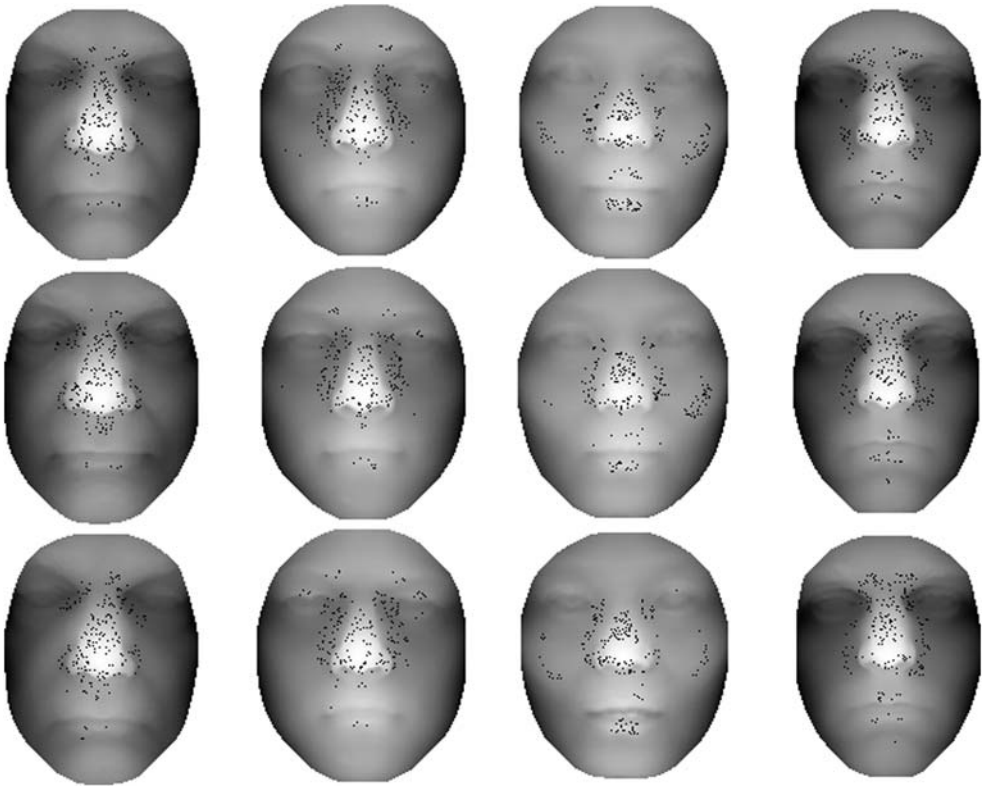


Fig. 1 Illustration of keypoint repeatability. Each *column* contains three range images of the same individual. Notice that the keypoints are repeatably identified at the same neighborhoods for the same individual. The height of range images is 160 mm which is why the keypoints appear very close

see that the keypoints are repeatably identified at the same locations for a given individual. Moreover, the keypoints vary between individuals because they have different facial shapes. For example, in the first column, the keypoints cluster mostly on the nose. In the second column, the keypoints cluster on the sides of the nose as well. In the third column, some keypoints are also detected on the cheek bones. The difference between keypoint locations on the faces of different individuals enhances the accuracy of the recognition phase.

Figure 2 shows the results of our keypoint repeatability experiment performed on the 3D faces of the FRGC v2 data (Phillips et al. 2005) i.e. 4007 three-dimensional scans of 466 individuals. The data was preprocessed for spike removal using neighborhood thresholding. Holes in the face data where then filled using cubic interpolation. Note that the FRGC consists of real data and therefore ground truth dense correspondence between the 3D faces is unavailable. Ground truth correspondence is necessary to calculate the error between the keypoints of different range images of the same person. Two approaches can be taken to overcome this problem. In the first approach which was also taken by Lowe (2004), data with ground truth is synthetically generated. This is done by adding synthetic changes like noise, rotation, scaling etc. which makes it possible to precisely calculate where each keypoint in an original image should appear in the transformed one. In the second approach, ground truth is approximated by an accurate registration algorithm. In this paper, we have adopted the latter approach because it is more realistic i.e. repeatability is calculated using real data. The 3D faces belonging to the same individual are automatically registered using a modified ICP algorithm (Mian et al. 2007) and the errors between the nearest neighbors of their keypoints (one from each face) are recorded. Figure 2 shows the cumulative percentage repeatability as a function of increasing distance. The repeatability reaches 86% for faces with neutral expression at an error of 4 mm. This is comparable to the repeatability of SIFT (Lowe 2004). As expected, the repeatability drops rapidly below 4 mm because this corresponds to the sampling distance (minimum distance) between the keypoints. For non-neutral expression faces, the repeatability at 4 mm drops to 75.6% because the 3D shape of the face changes with expressions. Note that these repeatability values are sufficient as the matching algorithm makes its decision on a subset of the total features (see Sect. 5).

The strengths of our keypoint detection algorithm can be summarized as follows. One, keypoint locations are repeatably identified in the range images of the same individual. Two, keypoints vary between different individuals. Three, keypoints are identified at locations where the shape variation is high (i.e. nonplanar and nonspherical regions). Four, the keypoint detection process also provides stable and repeatable local 3D coordinate frames for the computation of local features.

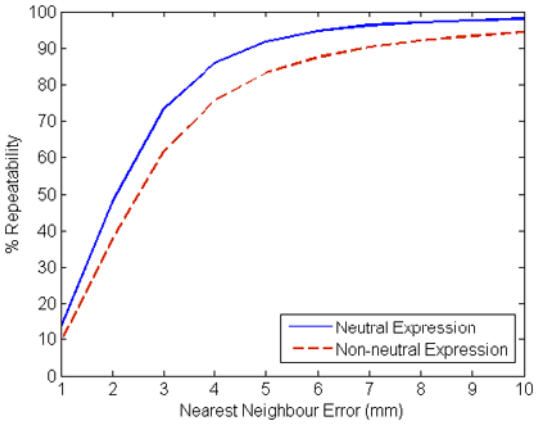


Fig. 2 Repeatability of keypoints

3 3D Feature Extraction

Once a keypoint has been detected, a local feature is extracted from its neighborhood $\mathbf{L}'$. The proposed local 3D feature is an extension of the tensor representation (Mian et al. 2006b, 2006e) which quantizes local surface patches of a 3D object into three-dimensional grids defined in locally derived coordinate bases. In Mian et al. (2006b), the local coordinate bases were derived from two points and their corresponding normals. However, in a later work (Mian et al. 2006d), the local coordinates were derived from a single point and some further invariant information i.e. the SIFT orientations or the location of the nose tip, in order to avoid the C_2^n combinatorial problem (Mian et al. 2006b). In this paper, we use the principal directions of the local surface patch $\mathbf{L}'$ as the 3D coordinates to calculate the feature. This avoids the C_2^n combinatorial problem (Mian et al. 2006b) without the knowledge of the nose tip (Mian et al. 2006d). Since the keypoints are detected such that there is no ambiguity in the principal directions of the neighboring surface patch, the derived 3D coordinate bases are stable and so are the features.

A surface is fitted to the points in $\mathbf{L}'$ using approximation as opposed to interpolation. This way the surface fitting is not sensitive to noise and outliers in the data. Each point in $\mathbf{L}'$ pulls the surface towards itself and a stiffness factor controls the flexibility of the surface. The surface is sampled on a uniform lattice. We use a 20×20 lattice in our experiments to extract the feature. Figure 3b shows a surface fitted to the neighborhood of a keypoint using a 20×20 lattice. In order to avoid the effects of boundaries that appear on the flanks of $\mathbf{L}'$, we crop a larger region first using r_2 (where $r_2 > r_1$) and fit a surface to it. This surface is then sampled on a bigger lattice and only the central 20×20 samples covering the r_1 region are concatenated to form a feature vector of dimension 400. For surface fitting, we used publicly available code (D'Erico 2006). However, any surface fitting algorithm that uses approximation can be used for this purpose.

Keeping a constant value for the threshold t_1 will result in different numbers of keypoints identified for different individuals. However, we have put an upper limit on the total number of local features that are calculated for a face in the gallery. This is important in order to avoid the recognition results being biased in favor of the gallery faces that have more local features. For every face in the gallery, a total of 200 feature vectors are calculated. The 200 keypoints are selected using a uniform random distribution. A dimension of 400 is quite large for a feature vector that describes a local surface. Fortunately, it is possible to compress these vectors by projecting them into a subspace defined by the eigenvectors of their largest eigenvalues using Principal Component Analysis (PCA). Let $F = [\mathbf{f}_1 \ldots \mathbf{f}_{200N}]$ (where N is the gallery size) be the $400 \times 200N$ matrix of all the feature vectors of all the faces in the gallery. Each column of F contains a feature vector of dimension 400. The mean of F or the mean feature vector is given by

$$\bar{\mathbf{f}} = \frac{1}{200N} \sum_{i}^{200N} \mathbf{f}_i. \tag{7}$$

The mean feature vector is subtracted from all features

$$\mathbf{f}'_i = \mathbf{f}_i - \bar{\mathbf{f}}. \tag{8}$$

The mean subtracted feature matrix becomes

$$F' = [\mathbf{f}'_1 \ldots \mathbf{f}'_{200N}]. \tag{9}$$

The covariance matrix of the mean subtracted feature vectors is given by

$$\mathbf{C} = F'(F')^T, \tag{10}$$

where $\mathbf{C}$ is a 400×400 matrix. The eigenvalues and eigenvectors of $\mathbf{C}$ are calculated using Singular Value Decomposition (SVD) as

$$\mathbf{USV}^T = \mathbf{C}, \tag{11}$$

where $\mathbf{U}$ is a 400×400 matrix of the eigenvectors sorted in decreasing order i.e. the eigenvector corresponding to the highest eigenvalue is in the first column. $\mathbf{S}$ is a diagonal matrix of the eigenvalues, also sorted in decreasing order. The dimension of the PCA subspace is governed by the amount of required accuracy (fidelity) in the projected space. This can be easily understood by plotting the ratio of the first k eigenvalues to the total eigenvalues given by

$$\psi = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{400} \lambda_i}, \tag{12}$$

where λ_i is the ith eigenvalue. Figure 4 shows a plot of the ratio ψ as a function of the number of eigenvalues k. The

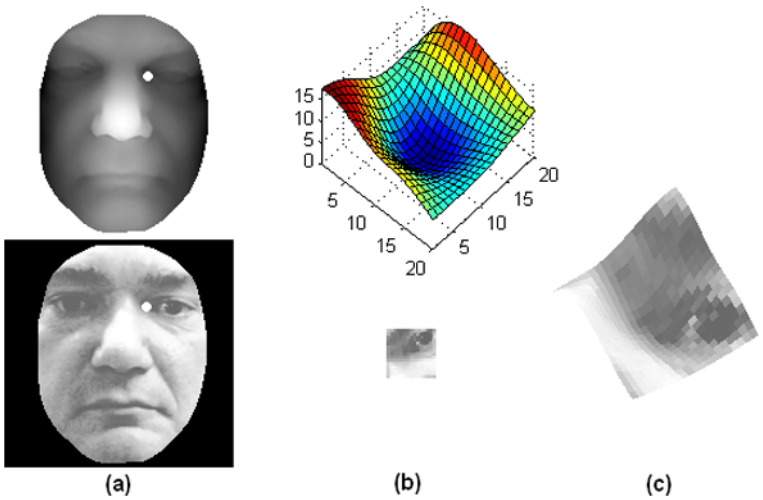


Fig. 3 **a** A keypoint displayed (in *white colour*) on a 3D face (*top*) and its corresponding texture map (*bottom*). **b** A local surface fitted to the neighborhood of the keypoint (*top*) using a 20×20 lattice. The corresponding texture map is also rotated (*bottom*) for alignment with the coordinates of the 3D surface. **c** The texture mapped on the local surface

curve reaches a value of 0.99 very quickly at only $k = 11$ which means that taking eigenvectors corresponding to only the highest 11 eigenvalues gives us 99% accuracy and a compression ratio of $\frac{(400-11)}{400} = 97.3\%$. This is not surprising given that all human faces have a similar topological structure and are roughly symmetric on either side of the nose. Repeating the keypoint detection with different values of parameters (r_1, t_1) as well as randomly selecting keypoints showed that the value of $k = 11$ is repeatable. The value of $k = 11$ has certain significance here. It shows that local shape variation in 3D faces can be represented with 99% accuracy by a vector of fairly small dimension. This is in contrast with 2D faces where local appearance variation is represented with only 60% accuracy by a vector of equal dimension (see Sect. 4.2). The first k eigenvectors are taken as

$$\mathbf{U}_k = \mathbf{U}_i, \tag{13}$$

where $i = 1, \ldots, k$ and $\mathbf{U}_k$ is a $400 \times k$ matrix of the first k eigenvectors. The mean subtracted feature matrix is projected to the eigenspace as

$$F^{\lambda} = (\mathbf{U}_k)^T F', \tag{14}$$

where F^{λ} is a $k \times 200N$ matrix of the 3D feature vectors of the gallery faces. F^{λ} is normalized so that its variance along each of the k dimensions is equal

$$F^{\lambda}_{rc} = \frac{F^{\lambda}_{rc}}{\lambda_r} \quad \text{where } r = 1, \ldots, k \text{ and } c = 1, \ldots, 200N. \tag{15}$$

In (15), r stands for the dimension or row number and c stands for the feature or column number. Finally, the feature vectors in F^{λ} (i.e. the columns) are normalized to unit magnitude and saved in a database along with $\overline{f}$ and $\mathbf{U}_k$ for online feature-based recognition of faces. Note that the representation of faces in the gallery is quite compact. Each face is represented by only 200 feature vectors each of dimension 11.

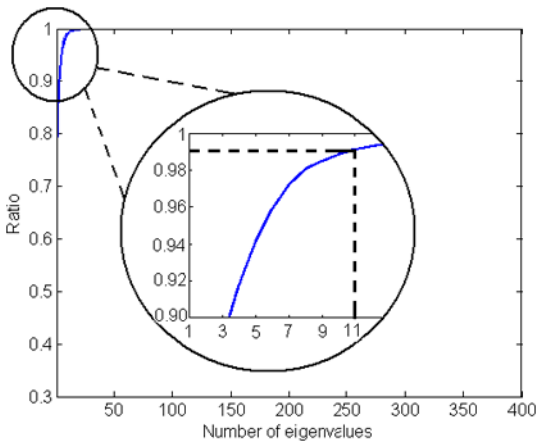


Fig. 4 A plot of the ratio ψ as a function of the number of eigenvalues k shows that the eigenvectors corresponding to the highest 11 eigenvalues give 99% accuracy

4 2D Feature Extraction and Fusion

4.1 2D Feature Extraction

In the 2D domain we used an off the shelf feature extraction algorithm. We chose SIFT (Scale Invariant Feature Transform) (Lowe 2004) for this purpose as our keypoint detection algorithm was inspired by the SIFT keypoint detection even though the two algorithms take completely different approaches to keypoint detection. In SIFT (Lowe 2004), a cascaded filtering approach is used to locate the keypoints which are stable over scale space. First, keypoint locations are detected as the scale space extrema in the Difference-of-Gaussian function convolved with the image. A thresh-

old is then applied to eliminate keypoints with low contrast and those 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 corresponding to the peaks in the histogram are used to assign one or more orientations to the keypoint.

For every orientation of a keypoint, the gradients of its local neighborhood are used to extract a feature (SIFT). SIFT is invariant to orientations because it is extracted after rotating the gradients relative to the keypoint orientation. 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 $4 \times 4 \times 8 = 128$ dimensional feature vector. To achieve robustness to illumination, the feature vector is normalized to unit magnitude and large gradient magnitudes are thresholded to a ceiling of 0.2 each and the vector is renormalized to unit magnitude once again. A detailed explanation of the SIFT keypoint detection and feature extraction is given by Lowe (2004).

4.2 Feature-level Fusion

It is believed that feature-level fusion can provide better results than score-level fusion (Jain et al. 2004). However, current research in the area of feature-level fusion is still in its infancy. Some of the challenges in feature-level fusion include the relative incompatibility of features and dimensionality problems. Incompatibilities can be in the dimensionality of the features, the variance of the features along each dimension and the location from which these features are extracted. In this section, we address these problems one by one. First, we standardize the keypoint locations so that both 2D and 3D features can be extracted from the same points. Second, we reduce the dimensionality of the 2D feature and normalize it so that both 2D and 3D features contribute equally to the resultant fused features.

The process described in Sect. 4.1 generally identifies 2D keypoints at locations that are different from the 3D keypoints described in Sect. 2. This is not a problem when both features are matched independently and fusion is performed at the score-level. However, for feature-level fusion, both the 2D and 3D features must be extracted at the same locations. There are two ways to do this. The first is to extract 3D features at points on the 3D face that correspond to the 2D keypoints on the 2D image. This approach was tested in our earlier work (Mian et al. 2006d) but the results indicated that the keypoint locations or orientations provided by the 2D features are not suitable for extracting 3D features.

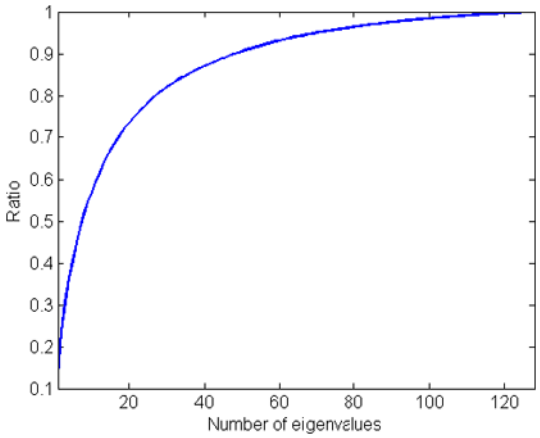
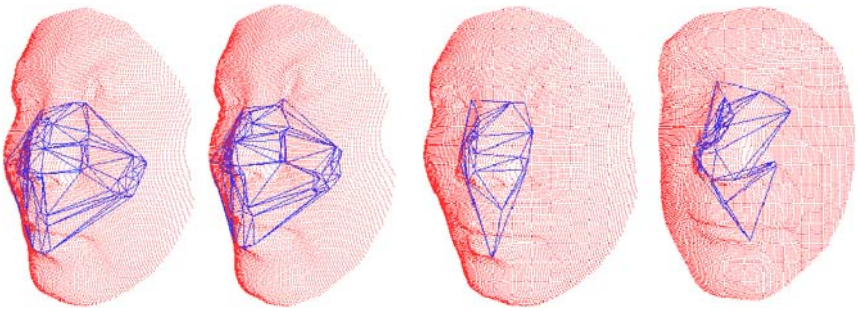


Fig. 5 A plot of the ratio ψ as a function of the number of eigenvalues k for the 128 dimensional SIFT features shows that the eigenvectors corresponding to the highest 64 eigenvalues give 95% accuracy

The second approach is tested in this paper whereby SIFT features are extracted at the 2D image locations which correspond to the keypoint locations and orientations on the 3D face. The 2D pixel values corresponding to the local region $\mathbf{L}$ are mapped on the point cloud and rotated using (4) in order to align them with the coordinates of the 3D feature. These pixels are then sampled on a uniform 20×20 grid so that the pixel-to-point correspondence is maintained to the local surface. Figure 3b shows a local 2D image patch after rotation and sampling. The same image patch is mapped onto its corresponding 3D surface in Fig. 3c.

Note that using the 3D keypoints for extracting SIFTs has an added advantage that the image scale can be normalized with respect to the absolute 3D scale. In other words, SIFTs can be extracted from the local 2D image patches at the same scale as that of the local surface. Once the 2D SIFT features are extracted from the same keypoints as the 3D ones, they are projected to the PCA subspace in a similar way as described in Sect. 3. To decide on the number of SIFT dimensions (eigenvectors) that must be used, the ratio ψ is plotted versus k in Fig. 5. Ideally, it is desirable to keep the dimensions of the 3D and 2D features equal so that the resultant feature and the matching process is not biased. However, this is not feasible as the two features have significantly different reconstruction accuracies at equal values of k (see Fig. 4 and Fig. 5). For example, at $k = 11$ the accuracy of SIFT is only 0.6 whereas that of the 3D features is 0.99. To strike a balance between accuracy and compression, we selected eigenvectors of SIFTs that correspond to the 64 highest eigenvalues (i.e. $k = 64$). The variance of the 64 dimensional vectors along each dimension is then normalized using (15) and the resultant vector is normalized to unit magnitude. Corresponding 2D and 3D feature vectors are concatenated and normalized to unit magnitude to form a multimodal feature vector.



(a) (b)

Fig. 6 **a** Correct match. **b** Incorrect match

5 Feature Matching

It is possible to speed up the matching process with the help of indexing or hashing (Mian et al. 2006e). However, these techniques are not included here because we are interested in matching a probe to every gallery face in order to generate a large number of impostor scores which are useful for drawing the Receiver Operating Characteristic (ROC) curves. Currently, an unoptimized implementation of our algorithm in MATLAB can perform 23 matches per second using a 3.2 GHz Pentium IV machine with 1 GB RAM.

During online recognition, features are extracted from the probe face using the same parameters as the gallery. Let $\mathbf{f}_p$ be a feature extracted from a probe face. This feature is first projected to the PCA subspace

$$\mathbf{f}_p^{\lambda} = (\mathbf{U}_k)^T(\mathbf{f}_p - \bar{\mathbf{f}}). \tag{16}$$

To calculate the similarity between a probe and a gallery face, their local features are matched using the following equation

$$e = \cos^{-1}(\mathbf{f}_p^{\lambda}(\mathbf{f}_g^{\lambda})^T), \tag{17}$$

where $\mathbf{f}_p^{\lambda}$ and $\mathbf{f}_g^{\lambda}$ correspond to the probe and gallery features in the PCA subspace, respectively. These features could be 2D, 3D or multimodal 2D-3D. The same matching algorithm (to the code level) is used for matching different types of features. If the two features are exactly equal, the value of e will be zero indicating a perfect match. However, in reality a finite error exists between the features extracted from the exact same locations on different images of the same face. For a given probe feature, the feature from the gallery face that has the minimum error with it is taken as its match. Once all the features are matched, the list of matching features is sorted according to e. If a gallery feature matches more than one probe feature, only the one with the minimum value of e is considered and the rest are removed from the list of matches. This allows for only one-to-one matches and the total number of matches m is different for every matching of probe-gallery faces.

The keypoints corresponding to the matching features on the probe face are projected on the xy-plane, meshed using Delaunay triangulation and projected back to the 3D space. This results in a 3D graph. The edges of this graph are used to construct a graph from the corresponding nodes (keypoints) of the gallery face using the list of matches. If the list of matches is correct i.e. the matching pairs of features correspond to the same location on the probe and gallery face, the two graphs are deemed to be similar (see Fig. 6). The similarity measure between the two graphs is

$$\gamma = \frac{1}{n_\varepsilon}\sum_{i}^{n_\varepsilon} |\varepsilon_{pi} - \varepsilon_{gi}| \tag{18}$$

where ε_{pi} and ε_{gi} are the lengths of the corresponding edges of the probe and gallery graphs, respectively. The value n_ε is the total number of edges. Equation (18) is an efficient way of measuring the spatial error between the matching pairs of probe and gallery features. The similarity γ is invariant to the facial pose because the edge lengths of the graphs will not vary if the graph is rotated or translated. Another similarity measure between the two faces is calculated as the mean Euclidean distance d between the nodes of the two graphs after least squared error minimization. Outlier nodes which have an error above a threshold are removed before calculating the mean error. This threshold is determined by the sampling distance of the faces used to find the keypoints in Sect. 2.

The matching algorithm described above results in four measures of similarity between the two faces i.e. the average error $\overline{e}$ between the features, the total number of matches m between the faces, the graph edge error γ and the graph node error d between the two graphs. Except for m, all other similarity measures have a negative polarity (i.e. a smaller value means a better similarity). A probe is matched with every face in the gallery resulting in four vectors $\mathbf{s}_q$ of similarity measures (where q corresponds to one of the four similarity measures namely $\overline{e}$, m, γ and d). The nth element of each vector corresponds to the similarity between the probe with

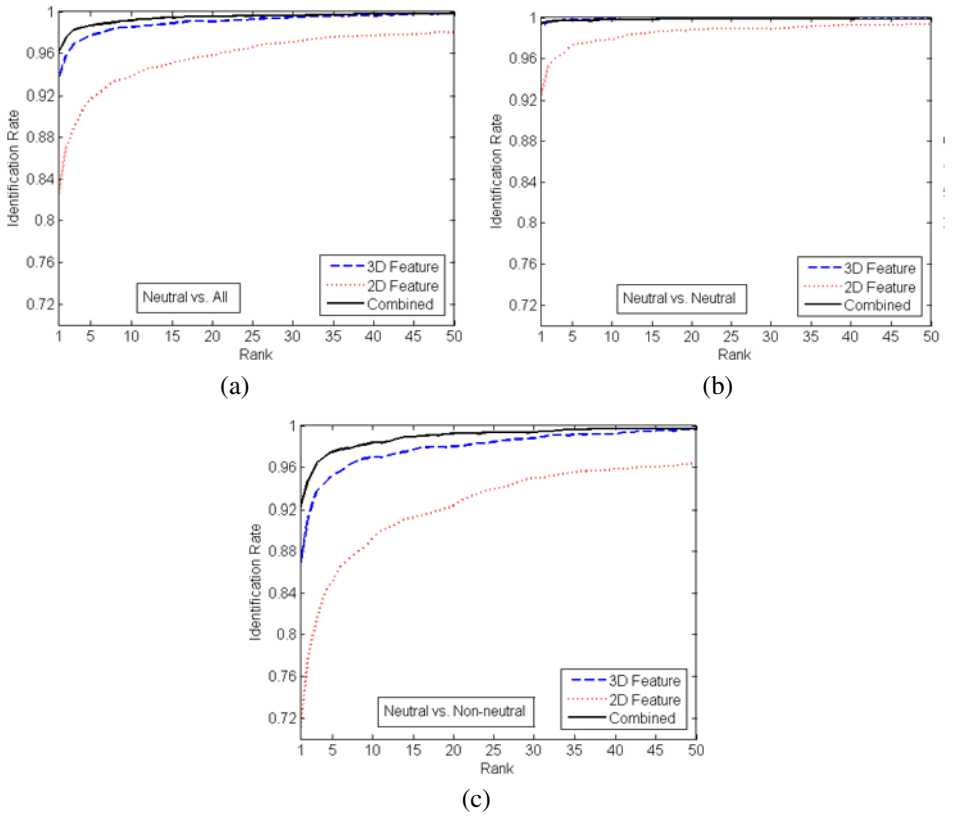


Fig. 7 Individual identification performance of the local features and their combined performance when fusion is performed at the score-level. The combined rank one identification rate for neutral versus all faces is 96.1%

the nth gallery face. Each vector is normalized on the scale of 0 to 1 using the min-max rule

$$\mathbf{s}'_q = \frac{\mathbf{s}_q - \min(\mathbf{s}_q)}{\max(\mathbf{s}_q - \min(\mathbf{s}_q)) - \min(\mathbf{s}_q - \min(\mathbf{s}_q))}, \tag{19}$$

where $\mathbf{s}'_q$ contains the normalized similarity measures. The operators $\min(\mathbf{s}_q)$ and $\max(\mathbf{s}_q)$ produce the minimum and maximum values of the vectors, respectively. The elements of $\mathbf{s}'_m$ (i.e. the number of matches) are subtracted from 1 in order to reverse their polarity. The overall similarity of the probe with the gallery faces is then calculated using a confidence weighted sum rule:

$$\mathbf{s} = \kappa_e \mathbf{s}'_e + \kappa_m(1 - \mathbf{s}'_m) + \kappa_\gamma \mathbf{s}'_\gamma + \kappa_d \mathbf{s}'_d, \tag{20}$$

where κ_q is the confidence in each individual similarity measure. Confidence measures can be calculated offline from the results obtained on training data or dynamically during online recognition as:

$$\kappa_q = \frac{\overline{s}_q - \min(\mathbf{s}_q)}{\overline{s}_q - \min_2(\mathbf{s}_q)}, \tag{21}$$

where $\overline{s}_q$ is the mean value of $\mathbf{s}_q$ and the operator $\min_2(\mathbf{s}_q)$ produces the second minimum value of the vector $\mathbf{s}_q$. Note that κ_m is calculated from $1 - \mathbf{s}'_m$. The gallery face which has the minimum value in the vector $\mathbf{s}$ is declared as the identity of the probe when the decision is to be made on the basis of individual 2D or 3D features or the fused multimodal 2D-3D features. In the case of score-level fusion, the similarity vectors resulting from individual 2D and 3D feature matching are normalized and fused using a weighted sum rule. The weights are calculated in a similar manner to (21). The resultant vector is then used to make the decision.

6 Results and Analysis

The FRGC v2 data consist of a training and a validation set. The validation set comprises 4007 three-dimensional scans of faces along with their texture maps. There are 466 subjects in the validation set. Minor pose variations and major expression and illumination variations exist in the database. A detailed description of the database can be found in Phillips et al. (2005). We selected one textured 3D face per individual under neutral expression to make a gallery of 466. The remaining faces ($4007 - 466 = 3541$) were treated as probes and were divided into two sets i.e. one with neutral expressions (1944 probes) and the other with non-neutral expressions (1597 probes).

Figure 7 shows our identification results. Note that it is not the aim of this paper to provide a true and unbiased comparison of the 3D and 2D features but to demonstrate their use for face recognition in the presence of expression

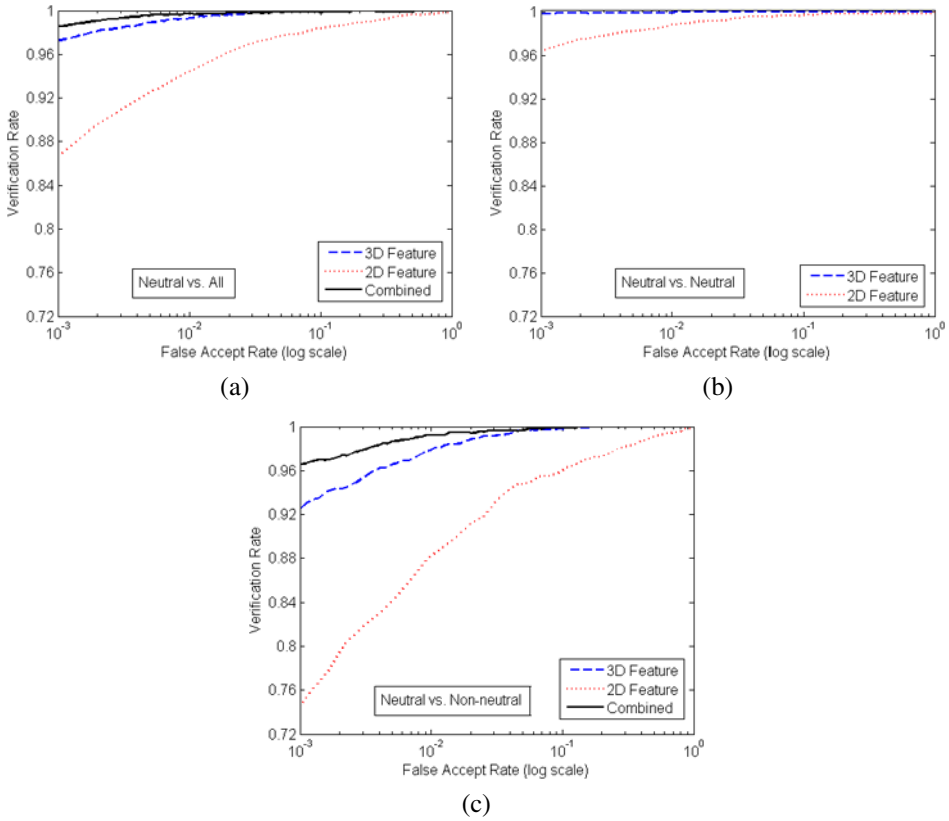


Fig. 8 Individual verification performance of the local features and their combined performance when fusion is performed at the score-level. The combined verification rate for neutral versus all faces is 98.6%

and illumination variations. The 3D local features achieved rank one identification rates of 99.0% and 86.7% for probes with neutral and non-neutral expressions, respectively. In the neutral expressions case, only two probes (out of a total of 1944) are above rank 11 and only one is above rank 17 i.e. a 100% recognition rate is achieved at rank 17. In the case of non-neutral expressions, the identification rate drops significantly however it should be kept in mind that 3D face recognition algorithms are generally more sensitive to expressions compared to 2D face recognition. For example, the face recognition rate in Lu et al. (2006) (using surface only) drops by 30% i.e. from 98% to 68%. However in our case, due to the use of local-features, the recognition rate drops by only 12.3%. Another point to note in the case of non-neutral expressions is that there is a steep rise in the recognition rate (i.e. 86.7% to 95%) from rank 1 to rank 5. This is an indication that it is possible to significantly improve the rank one recognition rate by fusing other features e.g. global. The combined (using 2D and 3D features) rank one identification rates are 99.4% and 92.1% for probes with neutral and non-neutral expressions, respectively.

Figure 8 shows the ROC curves of our algorithm. At 0.001 FAR (False Acceptance Rate), the 3D features achieved verification rates of 99.9% and 92.7%, respectively for probes with neutral and non-neutral expressions. In the neutral expression case, a 100% verification rate was achieved at 0.01 FAR. The combined verification rates are 99.9% and 96.6% for probes with neutral and non-neutral expressions, respectively. The combined results reported in Fig. 7 and Fig. 8 are for score-level fusion of the 3D and 2D features. Table 1 compares score and feature level fusion and recognition based on 3D features alone. Score-level fusion performs better than feature-level fusion which is contradictory to the claims of Jain et al. (2004) who argue that feature-level fusion is more powerful. Possible reasons for this anomaly are given in Sect. 7.

It is not the aim of this paper to report the most accurate results on the FRGC v2 data and we believe that better results can be obtained by combining our local 3D features with other features (e.g. global) in a multi-algorithm set up. However, to give some idea on the performance of our algorithm, we compare our results in Table 2 to others who used the FRGC v2 data set (Passalis et al. 2005; Maurer et al. 2005; Husken et al. 2005). The FRGC benchmark (verification rate at 0.001 FAR) is used for comparison. Table 2 shows that our algorithm has the highest 3D and multimodal verification rates in general. The performance of our local 3D feature-based face recognition especially stands out with 8% higher verification rate than its nearest competitor on the complete FRGC v2 data set (i.e. neutral versus all).

Table 1 Comparison of 3D feature-based face recognition with multimodal feature-based face recognition using score and feature-level fusion techniques

	Neutral vs. All		Neutral vs. Neutral		Neutral vs. Non-neutral	
	Identification Rate	Verification Rate	Identification Rate	Verification Rate	Identification Rate	Verification Rate
Score level fusion of 2D & 3D features	96.1%	98.6%	99.4%	99.9%	92.1%	96.6%
3D features alone	93.5%	97.4%	99.0%	99.9%	86.7%	92.7%
Feature level fusion of 2D & 3D features	93.1%	97.2%	97.8%	99.5%	87.4%	93.5%

Table 2 Comparison of verification rates at 0.001 FAR on the FRGC v2 data set

	Neutral vs. All		Neutral vs. Neutral		Neutral vs. Non-neutral	
	3D	Multimodal	3D	Multimodal	3D	Multimodal
This paper	97.4%	98.6%	99.9%	99.9%	92.7%	96.6%
Maurer et al. 2005	86.5%	95.8%	97.8%	99.2%	NA	NA
Husken et al. 2005	89.5%	97.3%	NA	NA	NA	NA
Passalis et al. 2005	85.1%	NA	94.9%	NA	79.4%	NA
FRGC baseline	45%	54%	NA	82%	40%	43%

7 Discussion

Not undermining the potential of feature-level fusion, there are four possible explanations for why this fusion technique did not even perform as well as the 3D features alone (see Table 1). The foremost reason is related to errors in the FRGC v2 data itself (Mian et al. 2007) as the 3D faces and their corresponding texture maps are not perfectly registered. Consequently, a keypoint detected on a 3D face sometimes corresponds to a different location on the texture map and hence the fused multimodal feature ends up distorted. The second reason is that given two sets of local features (2D + 3D), some features will be deteriorated in each set by variations due to noise, expressions and illumination etc. It is more likely that the deteriorated features from each set will belong to different keypoints and when the corresponding features of the two sets are fused, the number of deteriorated fused features will be greater than the those in either set. The third possible reason is that the same keypoints may not be suitable for extracting both 2D and 3D features. The fourth reason is that in score-level fusion, the two sets of features contain more information as they are extracted from different keypoints and therefore lead to better results. Ross and Govindarajan (2005) also report similar findings where score-level fusion performs better than feature-level fusion. Many other researchers have reported feature-level fusion techniques but did not compare their results to score-level fusion and one is left to speculate how their results will compare to score-level fusion.

Intuitively, one tends to agree with Jain et al. (2004) on the potential of feature-level fusion. However, there are many problems that need to be addressed before one could expect any improvement over score-level fusion. Our results show that score-level fusion currently remains the more robust (if not accurate) choice as it does not impose any restriction on the features in terms of size, variance, location and registration. Feature-level fusion can be considered as a global representation but in the feature domain as opposed to the spatial domain. Global features (in the spatial domain) are more sensitive to variations (Zhao et al. 2003) and by corollary one could imagine that feature-level fusion (global features in the feature domain) will also be more sensitive to variations. These arguments by no means rule out the advantages of interaction between multimodal features at the feature-level. Such interactions are beneficial as one feature could assist in the normalization or selection of another feature.

8 Conclusion

We presented a novel feature-based algorithm for the recognition of textured 3D faces. We proposed a keypoint detection algorithm which can repeatably identify locations on a face where shape variation is high. Moreover, a unique local 3D coordinate basis can be defined at each keypoint which allows for the extraction of highly descriptive 3D features. The 3D features were fused with existing 2D SIFT features at the score and feature-level and the performance of both fusion techniques was compared. We also proposed a graph-based feature matching algorithm and tested it on the largest publicly available corpus of textured 3D faces. Our algorithm has an Equal Error Rate (EER) of 0.45% (neutral

versus all case) and has the potential for further improvements if integrated with other features and/or classifiers in a multi-algorithm setup.

Acknowledgements We would like to thank the FRGC (Phillips et al. 2005) organizers for the face data, D'Erico for the surface fitting code, Lowe and El-Maraghi for the SIFT code. This research is sponsored by an ARC Discovery Grant DP0664228.

References

Al-Osaimi, F., Bennamoun, M., & Mian, A. S. (2006). Illumination normalization for color face images. In *International symposium on visual computing* (pp. 90–101).

Belhumeur, P., Hespanha, J., & Kriegman, D. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 711–720.

Besl, P. J., & McKay, N. D. (1992). 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), 239–256.

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

D'Erico, J. (2006). *Surface fitting using gridfit*. MATLAB Central File Exchange Select.

Gokberk, G., Salah, A., & Akarun, L. (2005). Rank-based decision fusion for 3d shape-based face recognition. In *International conference on audio and video-based biometric person authentication* (pp. 1019–1028).

Gonzalez, R. C., & Woods, R. E. (1992). *Digital image processing*. Reading: Addison–Wesley.

Huang, J., Heisele, B., & Blanz, V. (2003). Component-based face recognition with 3d morphable models. In *International conference on audio and video-based biometric person authentication*.

Husken, M., Brauckmann, M., Gehlen, S., & Malsburg, C. (2005), Strategies and benefits of fusion of 2d and 3d face recognition. In *IEEE workshop on FRGC experiments*.

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

Jones, M., & Viola, P. (2003). Face recognition using boosted local features. In *IEEE international conference on computer vision*.

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

Lu, X., Jain, A. K., & Colbry, D. (2006). Matching 2.5D scans to 3D models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(1), 31–43.

Maurer, T., Guigonis, D., Maslov, I., Pesenti, B., Tsaregorodtsev, A., West, D., & Medioni, G. (2005). Performance of geometrix ActiveID^{TM} 3D face recognition engine on the FRGC data. In *IEEE workshop on FRGC experiments*.

Mian, A. S., Bennamoun, M., & Owens, R. A. (2006a). 2D and 3D multimodal hybrid face recognition. In *European conference on computer vision* (pp. 344–355).

Mian, A. S., Bennamoun, M., & Owens, R. A. (2006b). A novel representation and feature matching algorithm for automatic pairwise registration of range images. *International Journal of Computer Vision*, *66*(1), 19–40.

Mian, A. S., Bennamoun, M., & Owens, R. A. (2006c). Automatic 3D face detection, normalization and recognition. In *3D data processing, visualization and transmission*.

Mian, A. S., Bennamoun, M., & Owens, R. A. (2006d). Face recognition using 2D and 3D multimodal local features. In *International symposium on visual computing* (pp. 860–870).

Mian, A. S., Bennamoun, M., & Owens, R. A. (2006e). Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1584–1601.

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

Passalis, G., Kakadiaris, I., Theoharis, T., Tederici, G., & Murtaza, N. (2005). Evaluation of 3D face recognition in the presence of facial expressions: an annotated deformable model approach. In *IEEE workshop on FRGC experiments*.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., & Worek, W. (2005). Overview of the face recognition grand challenge. In *IEEE computer vision and pattern recognition* (pp. 947–954).

Ross, A., & Govindarajan, R. (2005) Feature Level Fusion Using Hand and Face Biometrics. In *Biometric technology for human identification*. Bellingham: SPIE.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*, 71–86.

Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: a literature survey. In *ACM computing survey* (pp. 399–458).