Face Recognition using 2D and 3D Multimodal Local Features

Ajmal Mian, Mohammed Bennamoun, and Robyn Owens

School of Computer Science and Software Engineering,
The University of Western Australia,
35 Stirling Highway, Crawley, WA 6009, Australia
{ajmal, bennamou, robyn.owens}@csse.uwa.edu.au

Abstract. Machine recognition of faces is very challenging because it is an interclass recognition problem and the variation in faces is very low compared to other biometrics. Global features have been extensively used for face recognition however they are sensitive to variations caused by expressions, illumination, pose, occlusions and makeup. We present a novel 3D local feature for automatic face recognition which is robust to these variations. The 3D features are extracted by uniformly sampling local regions of the face in locally defined coordinate bases which makes them invariant to pose. The high descriptiveness of this feature makes it ideal for the challenging task of interclass recognition. In the 2D domain, we use the SIFT descriptor and fuse the results with the 3D approach at the score level. Experiments were performed using the FRGC v2.0 data and the achieved verification rates at 0.001 FAR were 98.5% and 86.0% for faces with neutral and non-neutral expressions respectively.

1 Introduction

The human face has emerged as one of the most promising biometrics due to its social acceptability and non-intrusiveness. It requires minimal or no cooperation from the subject making it ideal for surveillance and applications where customer satisfaction is important. However, face recognition is very challenging because it is an interclass recognition problem and the distinctiveness of face is quite low compared to other biometrics (e.g. fingerprints) [7]. 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 face recognition algorithms is given by Zhao et al. [17]. They also categorize face recognition algorithms into holistic, feature-based and hybrid matching algorithms. Holistic matching algorithms basically extract global features from the entire face. Eigenfaces [15] and Fisherfaces [1] 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 [13] which matches the 3D pointclouds of the eyes-forehead and the nose regions separately and fuse the results at the score level. Another example is the face recognition using local boosted features [8] which match 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. [6].

One limitation of holistic matching is that it requires accurate normalization of the faces according to pose, illumination 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. Replacing this manual process by an automatic feature 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 and 3D face recognition provides more accurate results than either of the individual modalities alone [3]. An up to date survey of 3D and multimodal face recognition is given by Bowyer et al. [3] who argue that 3D face recognition has the potential to overcome the limitations of its 2D counterpart however there is a need for better algorithms which are more tolerant to the variations mentioned above. Many 3D face recognition approaches are based on the ICP algorithm [2] or its modified versions. Advantages of ICP based approaches are that perfect normalization of the faces is not required and partial regions of faces can be matched with complete faces. The latter advantage has been exploited to avoid facial expressions [13] and to handle pose variations by matching 2.5D scans to complete face models [10]. 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 and thus rules out any possibility of indexing. Unless another algorithm and or modality is used to perform indexing, ICP based algorithms must perform a brute force matching thereby making the recognition time linear to the gallery size. Selecting expression insensitive regions of the face for matching 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.

In this paper, we present a face recognition algorithm using 2D and 3D multimodal local features. A novel 3D local feature is presented which is a modified version of the tensor representation [11] and extracts features in locally defined 3D coordinates. This makes the feature invariant to pose. Robustness to expressions is achieved by considering only the best predefined m matches. In the 2D domain, the SIFT descriptor [9] is used. SIFTs have mainly been used for pose and scale invariant 2D object recognition which is an intraclass recognition problem. To the best of our knowledge, their use for face recognition has not been thoroughly explored especially using the FRGC v2.0 data (Section 2). In this work, we use the SIFT descriptors for face recognition under illumination and expression variations. The results of the 2D and 3D local features are fused at the rank level using a confidence weighted sum rule. Preliminary experiments were performed on a *randomly selected* subset of the FRGC v2.0 data [14].

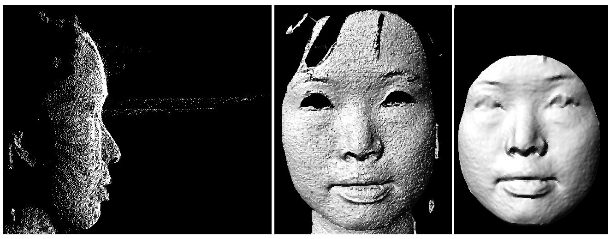


Fig. 1. (Left) A 3D pointcloud of a face shows spikes. (Center) The same face rendered as a shaded view to show noise. (Right) Shaded view after preprocessing.

2 Preprocessing the FRGC v2.0 Data

The FRGC v2.0 [14] defines a set of experiments and provides the largest available database for performing each experiment. Of these, only Experiment 3 is relevant to this paper i.e. matching 3D faces (shape and texture) to 3D faces (shape and texture). The FRGC v2.0 data for Experiment 3 consists of multiple 3D faces and their corresponding 2D faces of 466 individuals in the validation set. The database consists of frontal views with minor pose variations and major expression and illumination variations. The individuals are acquired from the shoulder level up (Fig. 2) and therefore a prior step of face detection is needed. Moreover, the 3D data is quite noisy and contains spikes and holes (Fig. 1).

Since preprocessing the data is not at the heart of this paper, we will describe it only briefly. For details, the reader is referred to [12]. A 3D face is automatically detected by locating the nose tip [12]. Next, the region of 3D face inside a sphere of radius r (where $r = 80$ mm) and centered at the nose tip is cropped. The corresponding pixels of the 2D face are also cropped at this stage. The spikes in the 3D face are then removed using a neighbourhood distance constraint and the holes are filled using cubic interpolation. The 3D faces are then median filtered to remove noise. Finally, the pose of the 3D face and its corresponding 2D coloured face is automatically corrected in an iterative algorithm based on the Hotelling transform [12]. The faces are also sampled on a uniform square grid at 1mm resolution during this process. The resultant faces have 161×161 pixels which is reasonable for 2D faces however in the case of 3D faces we delete alternate rows and columns to reduce their size to 80×80 pixels and 2mm resolution.

3 3D Local Features

Our novel local 3D feature is a variant of the tensor representation [11] which quantizes local surface patches of a 3D object into three-dimensional grids defined in locally derived coordinate bases. In [11], we derived the local coordinate basis from two points and their corresponding normals. In this paper, we define the coordinate basis using a single point in order to avoid the C_2^n (where n is the number of face data points) combinatorial problem [11]. We use a single point and its normal with some additional invariant information to define a local 3D coordinate basis (it is impossible to define a 3D coordinate basis using a single point and its normal alone). Two different types of invariant information were tested for this purpose. The first one was based on the orientation of the SIFT descriptor derived from the 2D face at the corresponding point (details

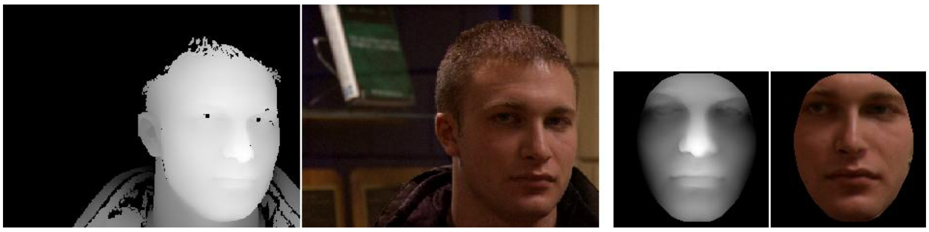


Fig. 2. Illustration of face detection and pose correction of a 3D face and its corresponding coloured texture map.

in Section 3.1). The second invariant information used was the location of the nose tip which was detected during the preprocessing (Section 2). Details of this approach are given in Section 3.2.

3.1 Deriving 3D Coordinates from SIFT Orientation and Normal

SIFTs (Scale Invariant Feature Transform) [9] are local descriptors computed at keypoints on a 2D image. These keypoints are extrema in the scale-space and are detected using the difference-of-Gaussian function. Each keypoint is assigned one or more orientations based on the local image gradient. The 2D coloured faces were converted to grayscale images and histogram equalization was used to reduce the effects of illumination before calculating SIFTs [9]. We use the orientations of each SIFT along with the normal of the keypoint calculated from the 3D data in order to define 3D local coordinate bases. The normal of the point, which is calculated by fitting a plane to the neighbouring points within a specified locality, makes the z-axis. The projection of the SIFT orientation on this plane defines the x-axis and the cross product of the z-axis with the x-axis defines the y-axis. Once the 3D basis is defined, the local 3D feature is computed as described in Section 3.3. Multiple orientations at a single point result in multiple 3D bases and hence multiple 3D local features. The downside of this approach is that since SIFTs are stable at only the keypoints, this restricts the number and location of the 3D features. This approach did not give satisfactory results (Fig. 7) indicating that the SIFT keypoint locations were not the best to calculate our 3D features. Another possible reason is that the SIFT orientations were not sufficiently stable to derive unique local 3D coordinate bases.

3.2 Deriving 3D Coordinates from Nose Tip and Normal

In this approach the 3D local coordinate basis at a point is derived using the normal of the point and the location of the nose tip. Since there is a single nose tip, this avoids the C_2^n problem [11] discussed above. Recall that the nose tip is automatically detected during the preprocessing stage (Section 2). The normal is again taken as the z-axis and the cross product of the z-axis with the vector from that point (where the 3D feature is to be calculated) to the nose tip defines the y-axis. The cross product of the y-axis with the z-axis defines the x-axis. Once the 3D basis is defined, the 3D feature is computed as described in Section 3.3. At this stage, we randomly select 300 locations on the face to extract the 3D local features. However, in future we plan to replace this random selection by a more reliable 3D keypoint identification algorithm. This approach gave better results (see Section 4) compared to the first approach probably because the the number and location of the 3D features are not tied up to the SIFTs and some of the 3D local features ended up being selected from regions which are more suitable for 3D features compared to the SIFT keypoint locations. Moreover, the coordinate bases defined using this approach were comparatively more stable.

3.3 3D Local Feature Extraction

Let $\mathbf{P}$ be a $3 \times n$ matrix of the x, y and z coordinates of the pointcloud of a 3D face given by Eqn. 1 (where n is the number of points).

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

The 3D local feature at a point $\mathbf{p}_i = [x_i \; y_i \; z_i]^\top$ is extracted as follows. First, all points $\mathbf{P}_l$ within a specified neighbourhood l of $\mathbf{p}_i$ are cropped and transformed to the local coordinate basis using Eqn. 2. Where $\mathbf{B}$ is the 3×3 matrix of the locally defined basis.

$$\mathbf{P}'_l = \mathbf{B}(\mathbf{P}_l - \mathbf{p}_i) \tag{2}$$

Eqn. 2 translates $\mathbf{P}_l$ so that $\mathbf{p}_i$ becomes the origin and rotates it so that it is aligned with the local coordinate basis. After rotation, the points in $\mathbf{P}'_l$ may no longer remain uniformly sampled. Therefore, $\mathbf{P}'_l$ is sampled again on a uniform grid in the xy-plane which is essentially the tangent plane used to calculate the normal of $\mathbf{p}_i$. This uniform sampling measures the distance of every sample point in $\mathbf{P}'_l$ to the tangent plane or a local invariant range image at that point. This range image is the 3D local feature at point $\mathbf{p}_i$ and is invariant to pose since it is defined in a local coordinate basis. The value of l decides the degree of locality of the feature and the resolution of the sampling decides the degree of granularity of the feature. Choosing a very high value for l makes the feature sensitive to facial expressions whereas choosing a very low value will make it less descriptive. Similarly, the sampling rate offers a trade off between accuracy and efficiency. We chose $l = 30mm$ and the sampling was done using a 30×30 on the basis of experiments performed on training data.

3.4 Matching 2D and 3D Local Features

It is possible to make the matching process much more efficient by using indexing or hashing. However, we performed a brute force matching in our initial experiments since our aim was to first demonstrate the effectiveness of these features for face recognition. To calculate the similarity between a gallery and probe face, their local features were matched using Euclidean distance. Features with minimum distance were considered as matches. Only one-to-one matches were established i.e. a feature from the gallery face was allowed to be a match to only one probe feature. The similarity score between the two faces was taken as the mean distance between the best predefined m matching pairs of features. This means that some matches were not considered allowing for variations caused in the data e.g. due to illumination and expressions.

Our novel local 3D features are highly descriptive and can find correct matches even in the challenging case of face recognition. Fig. 3 shows the five best matches

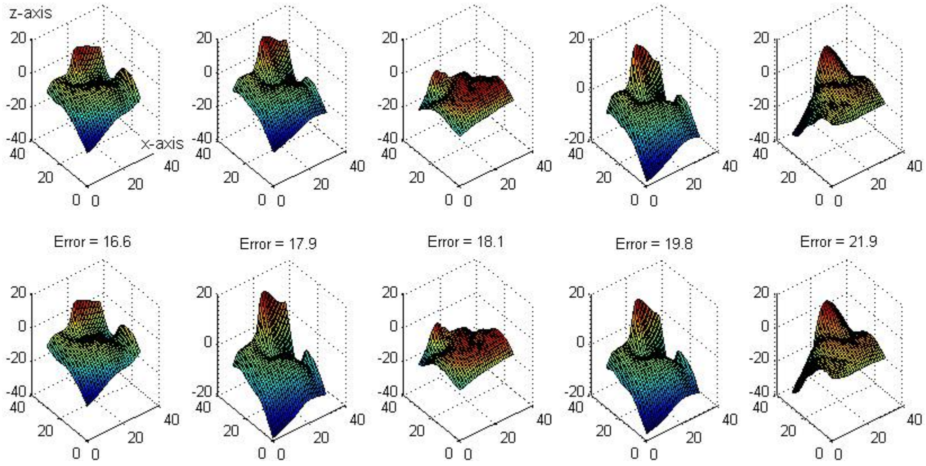


Fig. 3. Best five matches between the 3D local features of a probe (first row) with neutral expression and its correct identity in the gallery (second row). The 3D local features are rendered as 3D surfaces (the nose can easily be noticed as peaks in some of them). Each column shows a matching pair of 3D features with error written between them. Mean error was 39.4 for the best 100 matches in this case.

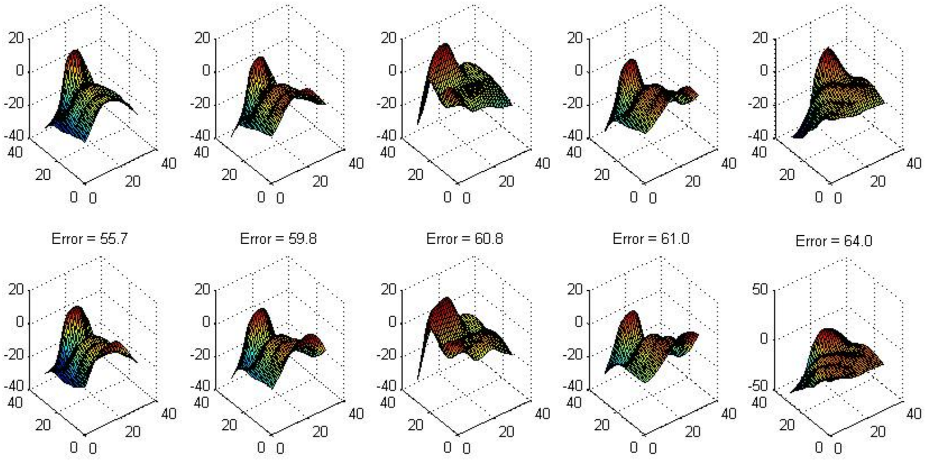


Fig. 4. Best five matches between different identities, under neutral expression, show higher errors compared to Fig. 3. Mean error was 92.8 for the best 100 matches.

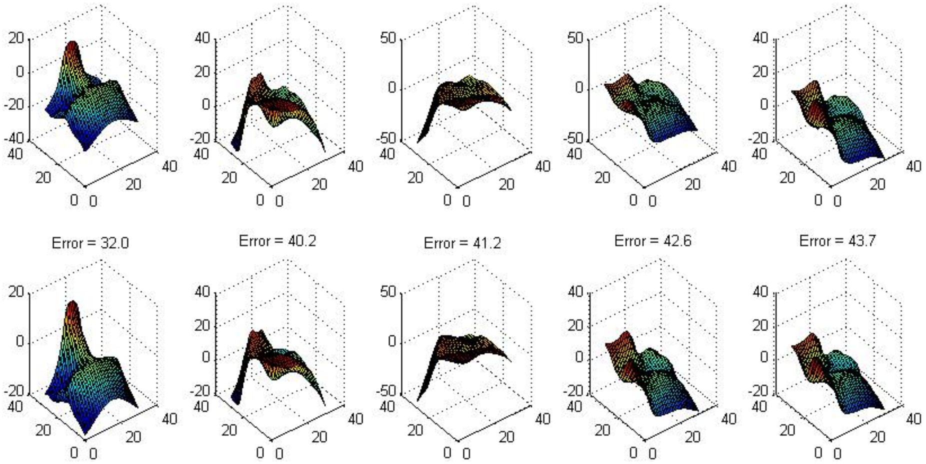


Fig. 5. Best five matches between the 3D local features of a probe (first row) with non-neutral expression and its correct identity in the gallery (second row). Notice that the errors are higher compared to the neutral expression case (Fig. 3) but are still much lower than in the case of an incorrect identity (Fig. 6). Mean error was 93.1 for the best 100 matches in this case.

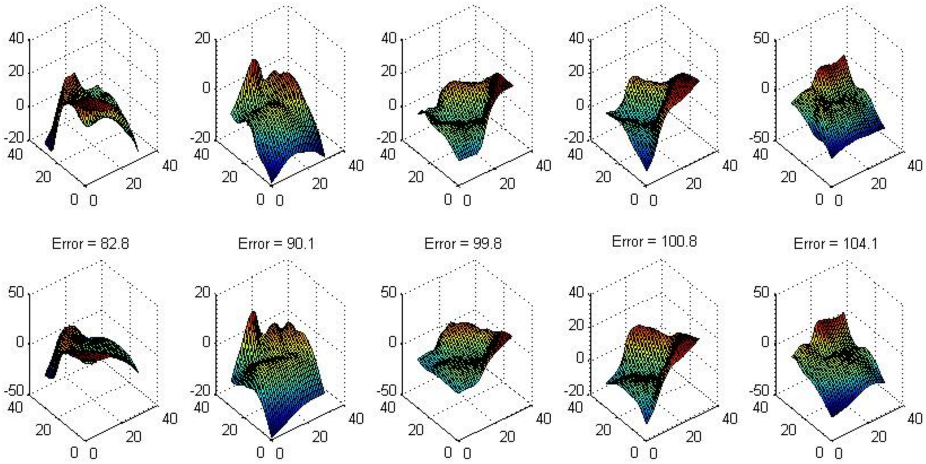
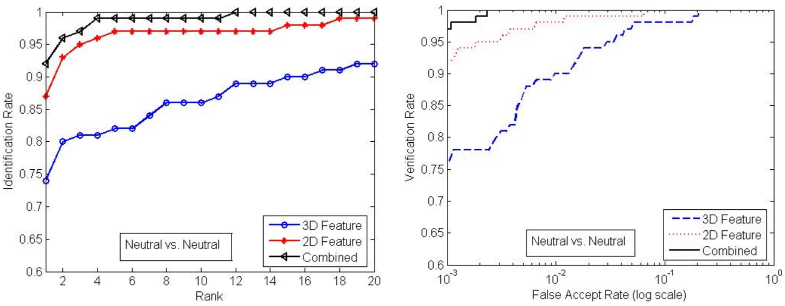


Fig. 6. Best five matches between different identities, under non-neutral expression, show higher errors compared to Fig. 5. Mean error was 156.1 for the best 100 matches.



(a) **(b)**

Fig. 7. (a) Identification and (b) verification performance when the 3D coordinate bases are derived from the point normals and SIFT orientations.

between a probe with neutral expression and its correct identity in the gallery whereas Fig. 4 shows the five best matches of the same probe with an incorrect identity. The first row corresponds to the 3D local features of the probe whereas the second row corresponds to those of the gallery. Each column represents a matching pair of features with error written between them. Notice that the errors are much lower in the case of the correct identity (Fig. 3). Under non-neutral expressions, the quality of the matches deteriorates however the correct identity still gives much lower error compared to the incorrect identity (Fig. 5 and 6).

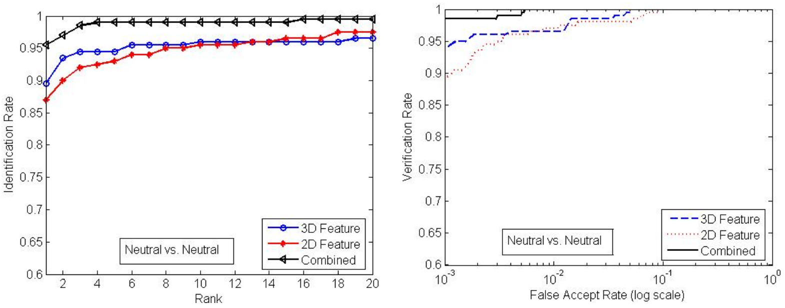
3.5 Fusion

The 2D and 3D local feature matching engines each results in a similarity matrix of size $N \times M$ (M is the gallery size and N is the number of probes tested) with negative polarity i.e. a lower value means higher similarity. The matrices are normalized using the min-max rule and then fused using a confidence weighted sum rule. Since each row of a similarity matrix corresponds to an independent recognition trial of a particular probe, the matrices are normalized row wise on the scale of 0 to 1. For each row (or recognition trial), the confidence value is calculated as the ratio of the difference between the best and mean similarity scores to the difference between the second best and mean similarity scores.

4 Results

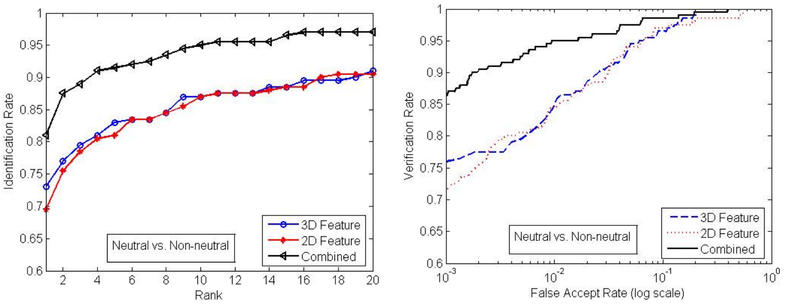
A single 3D face (neutral expression) per individual, along with its texture, was selected to build a gallery of 466 i.e. the maximum possible using the FRGC v2 data. This was to ensure a thorough and unbiased validation of our algorithm. For each experiment, 200 probes with neutral expression and another 200 with non-neutral expression were randomly selected to ensure that these samples are true representatives of their populations. The negligible difference between the 2D feature performance in Fig. 7 and 8 using different sets of randomly selected faces (neutral expression) supports our claim. All faces were preprocessed (Section 2) and their 2D (SIFT) and 3D local features were calculated and matched.

Fig. 7 shows our results when the local coordinate bases were derived from the SIFT orientations and the point normals. The 3D local features did not perform well in this case due to two possible reasons. One, the location of the SIFT keypoints are not suitable (in terms of descriptiveness) to the 3D local features. Two, the SIFT orientation does not provide stable local coordinates.



(a) **(b)**

Fig. 8. (a) Identification and (b) verification performance (under neutral expression) when the 3D coordinates are derived from the point normals and the nose tip location.



(a) **(b)**

Fig. 9. (a) Identification and (b) verification performance (under non-neutral expression) when the 3D coordinates are derived from the normals and the nose tip location.

In the next experiment, we derived the local coordinate basis from the point normals and the location of the nose tip. Fig. 8 and 9 show our results for probes with neutral and non-neutral expressions respectively. The results were very promising in this case as the 3D local features performed much better with individual identification rates of 89.5% and 73.0% for probes with neutral and non-neutral expressions respectively. The verification rates at 0.001 FAR for the same were 94.0% and 76.0% respectively. The 2D local features (SIFT) gave slightly lower but comparable performance to the 3D features. Note that it was not the aim of this paper to provide a true and unbiased comparison of these two features but to demonstrate their use for face recognition in the presence of illumination and expression variations. Fusion of the two features provides a significant improvement in performance with identification rates of 95.5% and 81.0% respectively for probes with neutral and non-neutral expressions. The verification rates at 0.001 FAR for the same were 98.5% and 86.0%.

5 Conclusion

We presented an automatic face recognition algorithm using 2D and 3D local features. We also presented a novel and highly descriptive 3D local feature and demonstrated its performance on a challenging interclass recognition problem i.e. face recognition. We effectively used the SIFT features for face recognition. By combining the 2D and 3D local features, we achieved a significant improvement in performance. Although, our preliminary results show some deterioration under non-neutral expressions, we believe that our 3D feature and recognition algorithm are promising and can be improved or used in conjunction with global features to give more accurate results. Moreover, the combined performance deterioration is significantly lower than that of the individual features. Our analysis show that the failures mainly occur because the 3D features are extracted at inappropriate locations. Therefore, in our future work, we would like to focus on identifying key locations for extracting the 3D local features.

6 Acknowledgment

We would like to thank the FRGC [14] organizers for providing data and D. Lowe for the SIFT code [9]. This research is sponsored by ARC Grant DP0664228.

References

1. P. Belhumeur, J. Hespanha and D. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection”, *IEEE PAMI*, vol. 19, pp. 711–720, 1997.
2. P. J. Besl and N. D. McKay, “Reconstruction of Real-world Objects via Simultaneous Registration and Robust Combination of Multiple Range Images,” *IEEE TPAMI*, vol. 14(2), pp. 239–256, 1992.
3. K. W. Bowyer, K. Chang and P. Flynn, “A Survey Of Approaches and Challenges in 3D and Multi-modal 3D + 2D Face Recognition,” *CVIU*, vol. 101, pp. 1–15, 2006.
4. C. S. Chua and R. Jarvis, “Point Signatures: A New Representation for 3D Object Recognition,” *IJCV*, vol. 25(1), pp. 63–85, 1997.
5. C. Chua, F. Han and Y. Ho, “3D Human Face Recognition Using Point Signatures,” *IEEE AMFG*, pp. 233–238, 2000.
6. J. Huang, B. Heisele and V. Blanz, “Component-based Face Recognition with 3D Morphable Models”, *AVBPA*, 2003.
7. A. K. Jain, A. Ross and S. Prabhakar, “An Introduction to Biometric Recognition,” *IEEE TCSVT*, vol. 14(1), pp. 4–20, 2004.
8. M. Jones and P. Viola, “Face Recognition using Boosted Local Features”, *IEEE ICCV*, 2003.
9. D. Lowe, “Distinctive Image Features from Scale-invariant Key Points”, *IJCV*, Vol. 60(2), pp. 91–110, 2004 (code available at http://www.cs.ubc.edu.ca/∼lowe/).
10. X. Lu, A. K. Jain and D. Colbry , “Matching 2.5D Scans to 3D Models,” *IEEE TPAMI*, Vol. 28(1), pp. 31-43, 2006.
11. A. S. Mian, M. Bennamoun and R. A. Owens, “A Novel Representation and Feature Matching Algorithm for Automatic Pairwise Registration of Range Images”, *IJCV*, vol. 66, pp. 19–40, 2006.
12. A. S. Mian, M. Bennamoun and R. A. Owens, “Automatic 3D Face Detection, Normalization and Recognition”, 3DPVT, 2006.
13. A. S. Mian, M. Bennamoun and R. A. Owens, “Region-based Matching for Robust 3D Face Recognition”, *BMVC*, vol. 1, pp. 199–208, 2005.
14. 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”, *IEEE CVPR*, 2005.
15. M. Turk and A. Pentland, “Eigenfaces for Recognition”, *JOCN*, Vol. 3, 1991.
16. C. Xu, Y. Wang, T. Tan and L. Quan, “Automatic 3D Face Recognition Combining Global Geometric Features with Local Shape Variation Information,” *IEEE ICPR*, pp. 308–313, 2004.
17. W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, “Face Recognition: A Literature Survey”, *ACM Computing Survey*, pp. 399-458, 2003.