Multibiometric human recognition using 3D ear and face features


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

ARTICLE INFO

Article history:
Received 25 July 2011
Received in revised form 17 September 2012
Accepted 20 September 2012
Available online 29 September 2012

Keywords:
3D face
Local 3D surface features
Multibiometric multimodal recognition
Score-level fusion
Feature-level fusion

ABSTRACT

We present automatic extraction of local 3D features (L3DF) from ear and face biometrics and their combination at the feature and score levels for robust identification. To the best of our knowledge, this paper is the first to present feature level fusion of 3D features extracted from ear and frontal face data. Scores from L3DF based matching are also fused with iterative closest point algorithm based matching using a weighted sum rule. We achieve identification and verification (at 0.001 FAR) rates of 99.0% and 99.4%, respectively, with neutral and 96.8% and 97.1% with non-neutral facial expressions on the largest public databases of 3D ear and face.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Traditional identity card and password based systems are increasingly exploited via theft or fakery [1,2]. Biometric recognition using human physiological (such as face, fingerprint, palmprint and iris) or behavioral (e.g., handwriting, gait and voice) characteristics is comparatively more robust to such frauds. In terms of acceptability, the face is considered the most promising biometric trait. Face data can be collected easily and non-intrusively. The face is also rich in distinct features. Very high recognition rates have been obtained for faces with neutral expression using three-dimensional (3D) or a combination of two-dimensional (2D) and 3D face images [3]. However, changes in facial expression significantly change the facial geometry [4] reducing the effectiveness of face recognition algorithms. To address this problem some researchers propose matching only rigid or semi-rigid regions/features of the face while others consider the face as non-rigid object and apply expression deformation modeling to 3D facial scans [5–10]. Some of them also apply multimodal approaches by combining multiple feature-based classifiers [11–16] or multiple instances of the same subjects [17]. However, most of these approaches fail if facial images are occluded with hair or ornaments. To address the challenges of both expressions and occlusions, multibiometric systems have recently been proposed where a decision is made based on a fusion or combination of different subsets of biometrics. It is relatively difficult to spoof multiple biometrics simultaneously.

Most of the popular multibiometric approaches fuse face data with other biometric modalities such as fingerprints [18], palm prints [19], hand geometry [20], gait [21], iris [22], voice [23] and most recently the ear (see Section 2.1). Among these alternative biometric traits, the ear has the advantage that it is co-located with the face and hence, respective data can easily be collected (with the same or similar sensor). Besides, changes of the shape of the ear over ages are less noticeable than those of the face. Recent literature reviews [24–26] in the area of ear biometrics reveal that encouraging recognition accuracies can be achieved using 2D [27–31] and 3D [32–34] ear data. Since the ear and face data have a very low correlation (which is a desirable criterion for any fusion approach), it is expected that the combination of the ear with the face biometrics will provide even better accuracy. Theocharis et al. [35] computed a correlation of only 0.16 between an ear and a face image using the Pearson correlation coefficient. They also illustrated that the ear–face fusion curve can reach a 100% recognition rate before rank-15 (i.e., the correct match is found at rank 15) whereas none of the modalities reached 100% accuracy before rank-20. This implies that the instances of failure to identify a subject using the two modalities are uncorrelated, and therefore, one modality can generally compensate for any grounding of the other.

Ear and face biometrics can be fused at different levels of the recognition process [36]. In this paper, we propose score and feature-level fusion approaches. We detect the face region of
interest from the frontal face images based on the position of the nose tip [14]. To detect the ear from the profile images, we use our previously developed ear detection technique using AdaBoost [37,38]. For score-level fusion, following a normalization step, face and ear local 3D features (L3DFs) are extracted and matched separately. Matching scores from the two modalities are then fused according to a weighted sum rule. For feature-level fusion, after extracting L3DFs from the ear and face data, we concatenate them based on their local shape similarity. We then match fused features according to the fused feature distance and using geometric consistency measures. Both these approaches are fully automatic and were found to be highly accurate and efficient when evaluated on the largest publicly available ear-face dataset using the Face Recognition Grand Challenge (FRGC) version 2.0 (FRGC v.2) [39] and the University of Notre Dame Biometrics Dataset Collection-J (UND–J) [40] databases. We had 315 and 311 probe (i.e., testing) images with neutral and non-neutral facial expressions respectively which were matched against 326 gallery images (i.e., enrolled off-line prior to testing) with neutral expression. We also evaluated the performance on a proprietary database collected in our laboratory comprising of face images with neutral and angry expressions and ear images with earphones from 56 subjects.

The remainder of the paper is organized as follows. Related work, motivations and contributions of this article are described in the next section. The data acquisition and feature extraction techniques are described in Section 3. The proposed approaches for score-level and feature-level fusions are described in Sections 4 and 5, respectively, along with their performance evaluations. Comparisons are made between proposed fusion approaches and with other approaches in Section 6. A conclusion is provided in Section 7.

### 2. Related work and contributions

#### 2.1. Related work

Multibiometric multimodal recognition with ear and face is a very recent research trend. Only very few approaches have been proposed using different levels of fusion [26]. Some of the most relevant approaches using score and feature levels of fusion are summarized in Table 1 and discussed below.

In score-level fusion, matching scores from different modalities are combined to make the recognition decision. Different fusion rules have been proposed. Kittler et al. [41] and Jain et al. [36] empirically demonstrated that the sum rule provides better results than other score fusion rules in a number of cases. Recently, Luciano and Krzyzak [42] demonstrated better results for the weighted sum rule.

There are only a few 2D approaches including the works of Luciano and Krzyzak [42,43] using score-level fusion of the ear and face biometrics. However, to the best of our knowledge, there are only two approaches using 3D data for score-level fusion of these two modalities. Considering their relevance to the research presented in this article, only 3D approaches are discussed as follows. Yan [44] combined ear and face at score-level using sum and interval fusion rules. On a database containing four ear and face images from each of the 174 subjects using earlier two images as galleries and the latter two images as probes, they obtained rank-one recognition rates of 93.1%, 97.7% and 100% for the ear, the face and the fusion respectively [45]. Theoharis et al. [35] extracted geometry images from 3D face and ear modalities and fitted annotated ear and face models using the iterative closest point (ICP) and a simulated annealing (SA) algorithm-based registration process. They applied the wavelet transform to

<p>| Table 1 | Multibiometric approaches using 2D and 3D ear and face data. |</p>
<table>
<thead>
<tr>
<th>Category</th>
<th>Source</th>
<th>Algorithms</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score-level fusion</td>
<td>Yan [44]</td>
<td>ICP and sum and interval fusion rules</td>
<td>Achieves high recognition accuracy (100%)</td>
<td>Involves using multiple images in the gallery dataset</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annotated model fitting, ICP, SA and wavelet transform</td>
<td>Achieves high recognition accuracy (99.7%)</td>
<td>Involves manual data extraction and substantial pre-processing</td>
</tr>
<tr>
<td></td>
<td>Theoharis et al. [35]</td>
<td></td>
<td></td>
<td>Tested on a smaller dataset of 174 subjects</td>
</tr>
<tr>
<td>Feature-level fusion</td>
<td>Chang et al. [46]</td>
<td>PCA</td>
<td>Simple</td>
<td>No separate test with non-neutral data</td>
</tr>
<tr>
<td></td>
<td>Xu et al. [47]</td>
<td>KFDA, minimum-distance classifier</td>
<td>Achieves good recognition accuracy (96.8%)</td>
<td>Verification rate is not reported</td>
</tr>
<tr>
<td></td>
<td>Pan et al. [48]</td>
<td>CCA, PCA, Minimum distance classifier</td>
<td>Achieves good recognition accuracy (97.4%)</td>
<td>Involves manual data extraction and substantial pre-processing</td>
</tr>
<tr>
<td></td>
<td>Xu and Mu [49]</td>
<td>KCCA, Minimum distance classifier</td>
<td>Achieves higher recognition accuracy (98.7%)</td>
<td>No separate test with non-neutral data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Verification rate is not reported</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Recognition time is not reported</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower recognition accuracy (90.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Uses 2D images</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Robustness to occlusion is not tested</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tested on a smaller dataset of 79 subjects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Based on 2D images</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tested on a very small dataset of 38 subjects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Based on 2D images</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tested on a very small dataset of 38 subjects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Based on 2D images</td>
</tr>
</tbody>
</table>
the extracted images to get individual feature vectors. The distance between the feature vectors of the gallery and the probe was weighted and then summed for fusion. They applied a weighted L1 (city-block) distance metric (defined as the shortest distance in unit steps along each axis between two points) to match the fused features where weights were empirically selected depending on the annotation of the face or ear model. The probe dataset for this experiment contained mostly faces with a neutral expression. As will be discussed in detail in Section 6, none of these papers quantified the verification performance, processing time and robustness to occlusions due to hair, ornaments or earphones.

In feature-level fusion, extracted features from different modalities are combined prior to matching. We are not aware of any 3D approaches using this fusion technique and only the following three 2D approaches are found in the literature. Chang et al. [46] proposed a principal component analysis (PCA) based multimodal approach of fusing face and ear biometrics. With a database of 197 2D images with no occlusion they achieved a recognition rate of 90.9%. Xu et al. [47] fused 2D features extracted from the profile face and ear using Kernel Fisher discriminant analysis (KFDA). They evaluated their technique using three 2D images (with rotations of $-5, 0$ and $+5$° around the vertical axis) per person in the gallery and six images per person in the probe dataset. They applied a minimum-distance classifier and the weighted sum rule (with weights 0.55 and 0.45 for face and ear respectively). Pan et al. [48] used canonical correlation analysis (CCA) to extract profile face and ear features. Prior to fusion, they reduced the dimension of the associated feature vector using PCA and used a minimum-distance based classifier for matching. While testing, they took three instances per gallery and two instances per probe for each subject. They obtained an identification rate of 97.4% using a fused feature vector of dimension 50. The same research group improved their result to 98.7% using kernel canonical correlation analysis (KCCA) [49].

2.2. Contributions

As reported in the literature review (Section 2.1), most of the score-level fusion techniques and all the feature-level fusion techniques use only 2D data for the fusion of ear and face modalities. However, 2D data are severely affected by changes in illumination, scale and pose variations that are common for public applications. Therefore, in this paper, we propose to use 3D data, which are commonly less sensitive to such variations. To the best of our knowledge, this is the first paper to present a feature-level fusion approach on 3D ear and face features.

Occlusions and deformations are very common in non-intrusive applications of ear-face biometrics. Most of the current ear–face multimodal approaches use global features that are affected by these variations. In this work, we use local 3D features (L3DFs) to represent both the ear and face data. L3DFs were first proposed by our research group [14] and used for object retrieval and face recognition [3]. The performance of these local features are compared with other similar features (such as local surface patch [32], Spin Image [50], sphere-spin-image [51] and iso-contours [52]) in our recent survey paper [26]. L3DFs are found to exhibit a high level of repeatability and very fast to compute and compare (23 matches per second on a 3.2 GHz Pentium IV machine using MATLAB [14]).

In comparison to other levels of fusion, score-level fusion has many benefits with respect to implementation and computation. It involves the processing of less data and consequently it is a faster and easier way when used to recognize people as noted by Jain et al. [36]. This has also been demonstrated in our previous work on score-level fusion [53] which has been extended and improved both in methodology (matching algorithm) and performance evaluation (on a new dataset) in this paper. Feature-level fusion, on the other hand, may preserve more discriminating features prior to matching and is intuitively believed to provide better accuracy. However, this has not been experimentally proven in the literature. Therefore, in this paper, we apply both fusion techniques using the same data and similarly extracted features and compare their recognition performance.

It is also interesting to note that all the existing feature-level fusion approaches use profile images for both the ear and face biometrics. Although it is easy and cost effective to use a single image for both modalities, a frontal face possesses more discriminating features than a profile face and hence may result in better recognition accuracy. To explore this, we propose to extract ear features from the profile image and face features from the frontal face images.

Our approaches are fully automatic all the way from data acquisition to recognition. We use our previously developed tools [54] to automatically extract a short list of candidates and select a minimal rectangular region from the whole dataset. Thus, we can minimize the cost when using ICP for improved accuracy, especially in the case of face images with non-neutral expressions.

The specific contributions of this article are as follows:

(a) Main contributions:

1. Two complete ear–face multimodal recognition systems are proposed based on 3D local features extracted fully automatically from the profile and frontal face images.

2. A novel feature-level fusion approach combining 3D ear and face features is presented that combines an ear feature and a face feature that have the minimum feature distance between them. Our results show that this is a worthwhile technique. It achieves accuracy comparable to that of the score-level fusion of the ear and face data without the need for the ICP-like computationally expensive algorithms in matching.

(b) Other contributions:

1. The largest possible multimodal dataset is constructed for the evaluation of the proposed fusion techniques comprising the publicly available profile (UND-J) and frontal (FRGC v.2) face databases.

2. A unique in-house multimodal database is developed by collecting frontal faces with three different expressions (smiling, angry and neutral) and profiles with and without earphones.

3. The development of similarity measures for matching multimodal gallery and probe data based on various geometric consistencies among the local 3D features of the ear and face which are significantly robust to non-neutral facial expressions.

3. Data acquisition, training and feature extraction

Multibiometric datasets collected and constructed to evaluate the performance of the proposed fusion techniques are discussed in this Section. A summary of the different characteristics of each dataset is provided in Table 2.

3.1. Gallery and probe datasets

We constructed a proprietary multimodal ear–face test dataset at the University of Western Australia (UWA) by acquiring 392 2D and their corresponding range images from 56 subjects of different sex, ages and ethnicities. The dataset was acquired in
the Computer Vision Research Laboratory at the School of Computer Science and Software Engineering, UWA using a Minolta Vivid 910 scanner in low-resolution mode in two different sessions with an average time lapse of four weeks. The experimental setup is shown in Fig. 1. For each of the subjects, we captured four frontal and three profile images. The frontal face images include two with neutral, one with smiling and one with angry expressions. One of the three profile images was taken when the subject wore an earphone. Images are divided into two groups: (a) gallery dataset in which ear images without earphones and face images with neutral expression collected in earlier sessions (off-line) will be enrolled prior to the matching phase and (b) probe dataset which includes the remaining images which will be matched against the images of the gallery dataset.

We also constructed a multimodal test dataset comprising ear images extracted from the Collection-J of the UND profile face database [34] and face images extracted from the Fall2003 and Spring2004 datasets of the FRGC v.2 frontal face database [55]. There are multiple 2D and their corresponding range images from 415 and 466 individuals in the UND-J and the FRGC v.2 databases respectively. We kept the earliest captured profile and frontal (with neutral expression) images of an individual in the gallery dataset and two of the remaining images in the probe dataset. For the gallery dataset, there are 326 individuals in the UND-J and the FRGC v.2 databases respectively. For the probe dataset, there are 311 and 315 individuals having face images with neutral and non-neutral (mostly smiling) expressions respectively in the FRGC v.2 database. Similarly, for the probe dataset, there are 311 and 315 individuals having face images with neutral and non-neutral (mostly smiling) expressions respectively in the FRGC v.2 database who have corresponding images in the UND-J database. To the best of our knowledge, this UND-FRGC test set is the largest publicly available ear–face dataset.

3.2. Training dataset

A small training dataset is constructed using ear images extracted from the first 100 subjects in the Collection-F of the UND profile database [40] and the corresponding face images in the FRGC v.2 face database. All the empirically determined parameters/thresholds used in the feature extraction (Section 3.4), matching (Sections 4.1.1 and 5.2) and the complementary weights used in Eqs. (2) and (5) are determined based on the training on this dataset. Details are summarized in Table 3.

3.3. Ear and face data extraction

In order to create the aforementioned datasets, ear and face data are extracted from the profile and frontal face images respectively. The ear region is detected from 2D profile images using our previously developed very accurate, fast and fully automatic approach. The ear detector is based on an optimized Cascaded AdaBoost algorithm and described in details in [37,38]. It consists of a set of 18 strong classifiers (in 18 stages) trained with thousands of weak classifiers. Eight different types of Haar-like features were used to generate the weak classifiers. The base size of the detector is $16 \times 24$ and a test image is scanned through

### Table 2
Multibiometric datasets used in this paper for performance evaluation.

<table>
<thead>
<tr>
<th>Database name</th>
<th>Gallery dataset</th>
<th>Probe dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image description</td>
<td>Number (each group)</td>
</tr>
<tr>
<td></td>
<td>Ear</td>
<td>Facial expression</td>
</tr>
<tr>
<td>UWA proprietary dataset</td>
<td>Without ear-phone</td>
<td>Neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>UND-FRGC (ear from UND-J and face from FRGC v.2 databases)</td>
<td>Mostly unoccluded and without any pose variation</td>
<td>Neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>326</td>
</tr>
<tr>
<td>Training dataset (UND-F and FRGC v.2 databases)</td>
<td>Mostly un-occluded and without any pose variation</td>
<td>Neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>
the detector in different sizes and locations. The trained classifiers in the detector are used in a cascaded manner, i.e., the classifier of one stage is only used when a sub-window in the test image is detected as positive (ear) by the classifier of its previous stage. A sub-window with a size of $16\times24$ or its multiple is accepted only when it passes through classifiers of all the stages. Our experiment with the detector on the UND-J dataset with 830 images from 415 subjects gives a detection rate of 99.9% (only one failure). It takes about 7.7 ms to detect the ear region in an image of size $640\times480$ using a C++ implementation on a Core 2 Quad 9550, 2.83 GHz machine.

During data collection, both the 2D textured color image and 3D scan of an object are captured simultaneously and available to the users in correspondence. Therefore, once the ear region is detected on the 2D profile image, the corresponding 3D data are then extracted from the coregistered 3D profile data as described in [38,56]. As shown in Fig. 1(a), a rectangular area of data points around the ear is extracted from the profile which sometimes includes some portion of the hair and the face. Therefore, the extracted data are pre-processed and normalized as follows. All the spikes and holes are removed by filtering the data. Triangulation is performed on the data points to remove long edges and subsequently disconnected points. Data are also shifted to their mean and then, uniformly sampled using a grid of size equal to the detection window plus 25 pixels in each direction.

The face region is initially detected from a $640\times480$ 2D frontal face image using Viola and Jones’ AdaBoost-based detector [57]. The position of the nose tip is then detected in the corresponding region of range image from which a sphere of radius of 80 mm (that sufficiently covers the frontal face area) is extracted as described in our previous work in [14]. We remove spikes from the extracted 3D face data, resample it on a uniform grid of 160 mm by 160 mm, remove holes (if any) using cubic interpolation and make pose correction using techniques described in [14].

### 3.4. Feature extraction

Local 3D features are extracted from 3D ear and face data. A number of distinctive 3D feature point locations (keypoints) are automatically selected on the 3D ear and 3D face region based on the asymmetrical variations in depth around them. As described in our previous works [3,38,54], these variations are determined by applying PCA to the neighborhood of a keypoint and then computing the difference between the eigenvalues along the first two principal axes. The number and locations of the keypoints are found to be different for the ear and the face images of different individuals and hence can be used as a digital signature for biometric purposes. It is also observed that these keypoints

<table>
<thead>
<tr>
<th>Parameter/ threshold</th>
<th>Definition</th>
<th>Impact of the value on the recognition performance</th>
<th>Chosen value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>Inner radius of the feature sphere within which data points are used to construct the 3D local features</td>
<td>Feature becomes more global and hence, more descriptive</td>
<td>Increases robustness to occlusion 15 mm</td>
</tr>
<tr>
<td>$r_s$</td>
<td>Seed resolution i.e., how close we choose a seed point</td>
<td>Less seed points</td>
<td>More seed points 2</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>Distance threshold limiting the distance between feature locations</td>
<td>More features will be included which may induce less significant matches</td>
<td>Reduces the number of matches 45</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>A threshold for rotation consistency</td>
<td>Allows considering matches having higher rotation variations in the calculation of rotation consistency</td>
<td>May discard potentially correct matches 10</td>
</tr>
<tr>
<td>$n_p$</td>
<td>The number of PCA eigenvectors determining dimension of the feature vector</td>
<td>Increases computational cost</td>
<td>May discard some important features 10</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Distance multiplier, the threshold to determine the distance consistency</td>
<td>Allows less consistent matches to be used in the construction of the classifier</td>
<td>More consistent matches to be used in the construction of the classifier 0.1</td>
</tr>
</tbody>
</table>

![Location of a local 3D feature on the range image of an ear](image1)

![Extracted 3D local surface feature](image2)

**Fig. 2.** Example of an extracted 3D local surface feature [54].
exhibit a high degree of repeatability in different images of the same individual [3,38].

Once a datapoint is determined as a keypoint, a sphere of radius $R$ is used to crop data to construct a 3D local feature. The value of $R$ is chosen relative to the average ear size. A larger radius will make the feature more descriptive, but less robust to occlusions due to hair, ear-rings, ear-phone or self-obstruction by the ear-ridges. The value of $R$ is empirically determined (on the training dataset) as 15 mm. The cropped datapoints are aligned on their principal axes and a uniformly sampled (with a resolution of 1 mm) 3D surface of $30 \times 30$ lattice is approximated using D’Errico’s surface fitting code [58]. In order to avoid boundary effects, an inner lattice of $20 \times 20$ is cropped from the bigger surface and converted to a 400-dimensional feature vector for that corresponding keypoint. An example of an extracted ear feature is illustrated in Fig. 2. More details can be found in [38].

The dimension of the feature vector is reduced by applying PCA to the local features of all the ears/faces of the gallery dataset. We performed a number of experiments using different numbers of PCA components ($n_p$), ranging from 10 to 70 with an interval of 10 on the training dataset (see Section 3.2). The results demonstrated insignificant performance differences for the variation in number of components and the best result was obtained for 10 components.

4. Unimodal matching and score-level fusion

The main steps in our multimodal recognition system using the score-level fusion are shown in the block diagram of Fig. 3. Each of the components is described in this section.

4.1. Matching technique

Ear images in the gallery and the probe datasets are matched using an L3DF-based coarse matching and an ICP-based fine matching techniques. A preliminary version of the approaches can be found in our previous works [38,54,59] on ear recognition. We adopt the same approaches with the same parameters for matching of the face data. The approaches are described briefly as follows.

4.1.1. L3DF-based matching

At first, root mean square (RMS) distance between a probe feature and all the gallery features are computed. Gallery features with a keypoint location at a predefined distance threshold $\bar{z}_1 = 45$ mm away from that of the probe feature are discarded. The remaining gallery features are sorted according to their RMS distance from the probe feature and the one with minimum distance is paired with the probe feature. If multiple probe features match the same gallery feature, the best match is retained.

The feature correspondence established above is improved in a second round of matching applying geometric (distance and rotation) consistency checks as explained in our previous work [38,59] and described briefly as follows.

Let $p'$ denote the 3D location of the keypoint of a feature in the probe dataset and $g'$ denote the location of the keypoint of the feature in the gallery dataset that is matched to that probe feature. Then for this match with locations $p', g'$, we find the count $n_m$ of how many other match locations $p'', g''$ (found in the first round between the gallery and the probe feature set) have probe and gallery distances that differ by less than the seed resolution (i.e., how close we chose a seed point) plus an error term proportional to the square root of the probe distance. The square root models the accumulation of error over a distance. Mathematically

$$\left| |p' - p''| - |g' - g''| \right| < \tau + \kappa \sqrt{|p' - p''|}$$

where $\tau$ is the seed resolution and $\kappa$ is a distance multiplier. The values of these parameters are empirically chosen as 2 mm and 0.1 respectively. A higher value of $\kappa$ will allow less consistent matches to be used in constructing the classifier.

After the determination of the count $n_m$ using Eq. (1) for all the matches that are found in the first round, we compute the proportion of distance-consistent matches ($\rho_d$) as the ratio of the maximum count (i.e., the maximum of $n_m$ across matches) and $n_t$, the total number of matches.
In the second round, we repeat the feature matching, allowing only those matches that are consistent with the most consistent match (i.e., matches with a feature distance less or equal to the distance in the match having maximum of \( m \)). We then compute the mean feature distance (\( \sigma_i \)) of the retained matches. An example of feature correspondence between a probe and gallery ear (after distance consistency check) is shown in Fig. 4. In this round, we also find the underlying rotation between the matched gallery-probe feature pair. For each of the matches selected above, we find the number (\( n_i \)) of other matches that have rotation angles within a threshold \( \rho = 10^\circ \). We compute the proportion of the consistent rotation (\( \rho_i \)) as the ratio of the maximum of \( n \) and \( n_i \). Finally, we align the keypoints of the matching probe features to those of the corresponding gallery features using ICP and record the ICP error as the keypoint distance measure (\( \sigma_i \)).

4.1.2. Candidate selection and ICP-based matching

According to the L3DF-based matching scores described in the above section, we select the best 40 gallery candidates (3D ear or face dataset of the subjects in the gallery database). These candidates are then aligned coarsely with the respective probe ear (or face) dataset using the transformation (distance and rotation) that maximally occur among the matched gallery-probe feature pairs. In order to align them finely, we employ a modified version of the ICP algorithm [14] on a minimal rectangular area of dataset containing only the matching features. We use the ICP error as a similarity measure (\( \sigma_i \)).

4.1.3. Score normalization and the final similarity measure

In order to make the final matching decision based on the fusion of ear and face modalities, we combine the following scores: (i) the mean feature distance of the matches retained by the second round (\( \sigma_i \)), (ii) the proportion of distance-consistent matches (\( \rho_i \)), (iii) the proportion of consistent rotations from the second round (\( \rho_i \)) and (iv) the ICP error (\( \sigma_i \)).

We normalize the scores or the values of each of the above similarity measures for a probe with the selected gallery candidates (computed during matching) on a 0 to 1 scale using the min-max rule. A confidence weight factor (\( \omega \)) is then computed for each of the similarity measures as the ratio between the minimum and second minimum of its values, taken relative to the mean. The final score (\( \eta \)) for a modality \( x \) (ear or face) is then computed as shown in Eq. (2) using a confidence weighted sum rule i.e., by summing the products of the scores and the corresponding weights

\[
\eta_x = \omega_x \sigma_i + \omega_y (1-\rho) + \omega_z (1-\rho) + \omega_o \sigma_i
\]

A list of the empirically evaluated parameters used in the extraction and matching of the ear and face features described above is provided in Table 3. The impact of the variation of their values in recognition performance and our choices are also summarized in the same table.

4.2. Fusion of scores

The matching scores from the ear and face data (\( \eta_e \) and \( \eta_f \)), respectively are fused using a weighted sum rule (similar to the one used in Eq. (2)) as follows:

\[
\epsilon = \omega_x \eta_e + \omega_f \eta_f
\]

where \( \omega_x \) and \( \omega_y \) are weights for the ear and the face modalities respectively.

Since L3DFs are more distinctive and reliable for the face data than the ear data, we augment the confidence weights (computed above) with some complementary weights during fusion. We empirically tried (on the multimodal training dataset) a range of values, and found that allocating nearly double weights to the face scores provides better results (see Fig. 8), e.g., weights of 2/3 and 1/3.

4.3. Performance evaluation

The recognition performance of our proposed score-level fusion approach is evaluated as follows. The fusion results with or without including the ICP scores in the final similarity measure are reported separately to demonstrate their individual contributions. The reported results are computed using the complementary ear and face weights of (0.35, 0.65) and (0.45, 0.55) for data with neutral and non-neutral facial expression respectively. Identification as well as verification (or authentication) rates are reported. The latter is an important indicator of the performance for some biometric systems where an identity is claimed by the users.

4.3.1. Results using L3DF-based measures only

The results computed without including the ICP scores (last term in Eq. (2)) i.e., only considering the L3DF-based scores are as follows. We obtain rank-one identification rates of 80.1% and 96.8% separately for the ear and face data respectively (rank-\( n \) means the right answer is in the top \( n \) matches) on the UND-FRGC test dataset with a neutral facial expression (see Section 3.1). The score-level fusion of these two modalities, improves the overall performance to 99.0% accuracy for rank-one identification. For the test data with non-neutral expressions, we obtain rank-one identification rates of 79.4%, 84.8% and 95.2% for the face, the ear and their score-level fusion respectively. The results are illustrated in Fig. 5.

The verification results at a false acceptance rate (FAR) of 0.001 on the UND-FRGC test obtained using only the L3DF-based similarity measures are illustrated with receiver operating characteristic (ROC) curves in Fig. 6. We obtain a verification rate of 99.4% for the fusion of ear and face with the neutral expression. For data with non-neutral expressions, the verification rate with face only is 84.8% which improves to 96.2% after fusion with the ear. Although expression does not affect 3D ear recognition, we observe a performance drop for the ear modality with face data with neutral and non-neutral expressions mostly due to the fact
that some of the testing 3D ear data corresponding to the face data with neutral expression are more challenging to recognize.

On the UWA proprietary dataset using local feature-based measures only, we obtain rank-one identification rates of 91.07% and 94.64% and verification rates (at 0.001 FAR) of 85.71% and 92.86% for the ear modality with earphones and without earphones, respectively. For the face with neutral expression, we obtain rank-one identification rate of 96.43% and verification rate of 98.21% which drop to 92.86% and 94.64% respectively, for the smiling expression. However, when we perform fusion of the two modalities, we obtain 100% identification and 98.21% verification accuracies which slightly reduce to 98.21% and 96.43% respectively, for the smiling expression. The values of the parameters and complementary weights used in these experiments were the same as used for the experiments with the UND-FRGC dataset.

4.3.2. Results including ICP scores

Considering the ICP scores from both the ear and face data during fusion, we obtain a slightly improved result for data with non-neutral facial expressions compared to those using L3DF-based measures only. The rank-one identification rates and verification rates at 0.001 FAR obtained for this approach on the UND-FRGC test set are reported in Table 4. It can be noted that although results of an individual modality improve with the use of ICP for neutral expression, the fusion result decreased slightly. This implies that we do not have to apply expensive post-processing (using the ICP algorithm) in applications where neutral facial expression can be ensured. ICP scores seem less reliable than L3DF-based matching scores in this case, and combining them is not worthwhile. We could instead use a different fusion approach such as the likelihood ratio-based fusion proposed by Nandakumar et al. [60]. However, this seems unlikely to improve the results much given the relative weakness of the ICP results.

Fig. 5. Identification results on the UND-FRGC test dataset for score-level fusion (without using ICP scores) of ear, and face with: (a) neutral expression and (b) non-neutral expressions.

![Identification results](image1)

![Identification results](image2)

Fig. 6. ROC curves for score-level fusion (without using ICP scores) of ear, and face from the UND-FRGC test dataset with: (a) neutral expression and (b) non-neutral expressions.

![ROC curves](image3)

![ROC curves](image4)
expressions do not reduce the recognition rates much compared to smiling expressions. However, in all cases we get significant improvements using the score-level fusion of the two modalities.

### 4.3.3. Effect of complementary weights

The variation of the recognition rates for different combinations of ear and face complementary weights (see Section 4.2) used for fusion with a weighted sum rule on the UND-FRGC test set with neutral facial expression is illustrated in Fig. 8. From the plot, we can see that the recognition rate reaches a peak for ear and face weights of 0.35 and 0.65 respectively. It then declines as greater weighting is given to the face. For data with non-neutral facial expressions, we obtain the best result with ear and face weights 0.45 and 0.55 respectively.

#### 4.3.4. Recognition speed

On a Core 2 Quad 9550, 2.83 GHz machine, a C++ implementation of our algorithms takes around 0.376 s for matching all the L3DFs of a probe image with those in all the gallery images and 1.36 s to match a gallery-probe pair using ICP.

### 5. Feature-level fusion

In this section, we describe our proposed approach for fusing local 3D features obtained from ear and face data. Fig. 9 shows a block diagram of the approach. The ear and face detection, 3D data normalization and feature extraction are performed as described in Section 3.3. The construction of the fused feature-set and their matching techniques are discussed below.
5.1. Fusion of L3DFs from ear and face

An individual’s ear and face have clearly different global shapes. However, as illustrated in Fig. 10, there exist some similarities among the local features extracted from these modalities. For example, there are some face and ear local features that form similar ridges. These similarities are generally incidental and not due to any particular relationship between an individual’s face and ear. Nevertheless, these similarities can be used to establish correspondences between ear and face local features in order to fuse these two modalities in a somewhat repeatable way. In particular, if a probe and a gallery have similar sets of ear and face features, they are likely to be fused in a similar way, leading to similar sets of fused features for the probe and the gallery.

Considering the above relationship among the corresponding ear and face features, for each of the ear features, we compute the RMS distance between that feature and all the face features of the same individual and sort them according to that distance. The face feature with minimum distance is paired up with the corresponding ear feature. We do the same for all ear features. Fusion can also be performed as face–ear combination by pairing a face feature with the most similar ear feature. We consider both combinations because the number of ear and face features are not always the same and one ear or face feature corresponds sometimes to multiple similar features. However, for uniformity of representation, we keep the ear features on the left side and the corresponding face features on the right side of the fused feature vector. As illustrated in Fig. 11, we also keep the two combinations as two halves of the fused feature-set.

The cumulative percentage of repeatability of the fused features in different images of the same person is shown in Fig. 12 for 10 subjects. While the repeatability is not ideal, one of the strengths of our fusion method is that it tends to work well (as reported in Section 5.3) even when features are not matched in a corresponding way. This is because ear and face features are matched when their feature-distance is small, and if, for example, two ear features are close then their corresponding two face features must also have a relatively small feature-distance due to the triangle inequality. Consider two ear features from corresponding parts of the ear on the probe and gallery images, \( E_p \) and \( E_g \). Then the distance \( |E_p - E_g| \) should be relatively small. Now suppose these are combined with face features \( F_p \) and \( F_g \). Then, even if the two face features are not from corresponding parts of the face, by the triangle equality we have

\[
|F_p - F_g| \leq |F_p - E_p| + |E_p - E_g| + |E_g - F_g|
\]

Since \( E_p \) and \( E_g \) are chosen to minimize \( |F_p - E_p| \) and \( |E_g - F_g| \) in the above inequality, overall this upper bound should be relatively small (particularly, as we reduce the feature dimension).

Each entry of the fused feature-set has 800 columns as each of the ear or face features is a vector of dimension 400. The number of rows of the fused feature-set depends on the number of ear features (n) and that of face features (m) (see Fig. 11). In order to
reduce feature dimension, we apply PCA to the fused feature vectors similar to the case of score-level fusion (Section 3.4). The number of selected PCA components \( n_p \) was empirically chosen (using training data) as 10.

### 5.2. Matching the fused features

In order to match a fused probe feature vector with a fused gallery feature vector, we apply a two-level matching technique similar to the one used for the ear or face features in the case of score-level fusion (see Section 4.1.1). However, the similarity measures are now computed with the constraint that the thresholds or limits are satisfied for both the ear and face features. For example, during the first stage of matching, we only allow fused gallery vectors whose both the ear and face features are within a distance threshold \( \lambda_1 \) away from the corresponding ear and face features in the probe vector. As described in Algorithm 1, we use five different similarity measures: mean feature distance of the first and the second stage \( \delta_1 \) and \( \delta_2 \) respectively, proportion of consistent rotations of the second stage \( \delta_3 \), proportion of distance-consistent matches \( \delta_4 \) and keypoint distance measure \( \delta_5 \).

#### Algorithm 1. Matching a multimodal probe with a multimodal gallery.

**input**: probe feature \( p \), gallery feature \( g \), distance thresholds \( \lambda_1 \), angle threshold \( \lambda_2 \), seed resolution \( \tau \), distance constant \( k \), minimum number of match \( m \)

**output** Similarity measures: \( \delta_1, \delta_2, \delta_3, \delta_4, \delta_5 \)

**Distance check**;
for Each feature in \( p \) and all features in \( g \) do
Discard gallery features with distance from the probe feature location > \( \lambda_1 \) for both ear and face features;
Pair the probe feature with the closest gallery feature, by both ear and face feature distance;
Count the number of matches \( n_t \) and discard the gallery if there is less than \( m \) feature pairs;
end

Compute the mean of the feature distance of the matching feature pairs \( \delta_1 \);

**Distance consistency check**;
for Each of the matching feature pairs do
find the number \( n_m \) of other matches that satisfy Eq. (1);
end
Find the match with maximum \( n_m \) and record it as \( T \);
Compute proportion of distance-consistent matches \( \delta_3 = \max(n_m)/n_t \);
**Second stage of matching**;
for All the gallery features do
Repeat step 1 (first for loop) but do not allow matches which are inconsistent with \( T \);
end
Compute the mean of the feature distance of the matching feature pairs \( \delta_4 \);
Discard the gallery if there is less than \( m \) feature pairs;

**Calculating proportion of rotation consistency measure**;
for Each of the selected matching pairs do
Find the number \( n_r \) of other matches that have rotation angles within \( \lambda_2 \);
Compute proportion of consistent rotation \( \delta_4 = \max(n_r)/n_r \);
end

---

**Fig. 11.** The block diagram of the feature-set constructed by the fusion of ear and face L3DFs. Each ear feature (e.g., \( E_1 \)) is concatenated with the corresponding best matched face feature (in this example \( F_{c1} \)). Similarly, each face feature (e.g., \( F_1 \)) is concatenated with its corresponding best matched ear feature (in this example \( E_{c1} \)). Thus, subscripts \( cn \) and \( cm \) indicate the index number of the closest feature with respect to its left or right feature for ear-face and face-ear combination, respectively.

**Fig. 12.** Repeatability of fused features in the gallery and probe images of 10 individuals.
Calculating keypoint distance measure:
Align the keypoints of the matching probe features to those of the corresponding gallery features using ICP;
Record the ICP error as keypoint distance measure ($\delta_b$);

We perform the above feature-based matching separately for both combinations (Fig. 11) of the fused feature-set. We then compute the mean of the first two similarity measures resulting from both combinations as $\delta_1$ and $\delta_2$ and retain other measures for the computation of the final similarity measure. All the similarity measures are normalized on a scale of 0 to 1 using the min–max rule. The confidence weighting factors $\eta_k$ (where $k=1$–5) are then computed as the ratio between the minimum and second minimum values of the corresponding similarity measures, taken relative to the mean (similar to the approach described in Section 4.1.3). The final similarity measure ($\epsilon_f$) is computed as a weighted sum of all these normalized similarity measures, with double weights to those obtained from the fusion with respect to the face feature, and given by Eq. (5) where subscript 'e' and 'f' are used to indicate that the corresponding similarity measure is computed from the ear–face and the face–ear combination of the features respectively:

$$
\epsilon_f = \eta_1 \delta_1 + 2\eta_2 \delta_2 + \eta_3 \delta_3 + 2\eta_4 \delta_4 + \eta_5 \delta_5 + 2\eta_6 \delta_6
$$

Note that the confidence weighting factors $\eta_k$ used in the above equation represent the confidence of individual similarity measure and are dynamically calculated at run time. It is possible that some of the weights are zeros in which case those particular similarity measures will not contribute to the overall score. The relative performance of the different similarity measures are illustrated in Fig. 14 and explained in Section 5.3.3.

5.3. Performance evaluation

The performance of the proposed feature-level fusion approach is evaluated on the same dataset as the one used for the score-level fusion (for a fair comparison) and discussed next.

5.3.1. Identification results

As illustrated in Fig. 13(a), for the feature-level fusion of ear and face features on the UND-FRGC test dataset with neutral facial expression, we obtain a rank-one identification rate of 98.4%. On the same dataset with non-neutral expressions, we obtain identification rates of 94.9%, 97.1% and 97.8% at rank-one, rank-two and rank-three respectively.

5.3.2. Verification results

For the data with neutral facial expression on the UND-FRGC test dataset, we obtain a verification rate of 99.0% at an FAR of 0.001. In the case of non-neutral expressions, the accuracy reduces to 96.8%. The results are illustrated in Fig. 13(b).

5.3.3. Performance of different similarity measures

The performance of different similarity measures used in feature-level fusion for face data with neutral expression is shown in Fig. 14. Similarity measures are named in short for the computation of the final similarity measure ($\epsilon_f$) and are dynamically calculated at run time. It is possible that some of the weights are zeros in which case those particular similarity measures will not contribute to the overall score. The relative performance of the different similarity measures are illustrated in Fig. 14 and explained in Section 5.3.3.

6. Comparative study

In this section, at first we discuss the comparison between our score-level and feature-level fusion approaches. We then provide
a comparison of our approach with other 3D ear–face multimodal approaches. A summary of the comparative analysis is provided in Table 6.

Table 6
Summary of the comparison of our fusion approaches with others. ('Id.' stands for identification and 'Ver.' stands for verification and the corresponding rates are measured at rank-one and at 0.001 FAR respectively.)

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Performance criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Id. rate (%)</td>
</tr>
<tr>
<td>Yan [44]</td>
<td>100</td>
</tr>
<tr>
<td>Theocharis et al. [35]</td>
<td>99.7</td>
</tr>
<tr>
<td>Score-level fusion of L3DFs (this paper)</td>
<td>99.0</td>
</tr>
<tr>
<td>Feature-level fusion of L3DFs (this paper)</td>
<td>98.4</td>
</tr>
</tbody>
</table>

6.1. Comparison between the proposed fusion approaches

The proposed score-level fusion approach achieves a better accuracy than the feature-level fusion approach for the multimodal recognition with the ear and face biometrics. Our previous work [3] with 2D scale invariant feature transform (SIFT) and 3D local features for multimodal face recognition also demonstrated similar results. This is mostly due to the difficulty in fusing local features in a repeatable way. As illustrated in Fig. 12, we only obtained a 40% repeatability of the ear-face fused feature vector even with 10 mm nearest neighbor error. However, the results reported in Section 5.3 express one strength of our feature-level fusion technique that it performs well even when features are fused differently between a probe and gallery. If features could be fused in the same way, the technique may outperform score-level fusion.

6.2. Comparison with other approaches

Yan [44] and Theocharis et al. [35] used a concise manual extraction and/or a substantial preprocessing step (e.g., using the snake algorithm and removing the face or the neck area using a skin detection algorithm) of the ear data which might have contributed to their higher identification accuracies. In contrast, our approach uses a fully automatic extraction technique that extracts a rectangular area around the ear, sometimes including extra regions with hair and skin with a minimal pre-processing to remove holes or spikes from the extracted 3D ear data.

The performance of the approach in [44] was evaluated on a smaller database from only 174 subjects each having two ear images and two face images in the gallery and the probe datasets. Faltiemier et al. [61] experimentally demonstrated that the multi-instance approach performs better than the single-instance approach, however, with a penalty of additional computation. We use a larger database without any multi-instance in the gallery or the probe datasets (Section 3.1).

None of the above approaches performed separate experiments with face data under non-neutral expressions that severely affect the performance. Theocharis et al. [35] used a subset of the FRGC v2 dataset which includes faces with non-neutral expressions, but the authors did not mention how many of their selected faces were with non-neutral expressions. On a larger dataset but with multi-instance gallery and probe datasets collected from the FRGC v2 database, Mian et al. [3] obtained 86.7% and 92.7% identification and verification rates respectively using face L3DFs involving non-neutral expressions. In this paper, we obtain better results (95.2% and 96.2% respectively) by fusing scores from ear L3DFs and face L3DFs (without considering ICP scores).

The authors of [35] and [44] did not report verification performance of their approaches which is critical to some biometric applications (see Section 4.3). Our approaches show high verification accuracy for face data with both neutral and non-neutral expressions.

The matching time is also not reported in either of these two approaches. However, since Yan [44] used ICP on the whole dataset, our technique with local features is expected to be faster than that approach. The final matching in [35] is expected to be faster, since they used a weighted L1 distance metric to compare wavelet coefficient extracted from ear and face data. However, their registration and deformable model fitting steps are computationally expensive.

7. Conclusion

In this paper, two multimodal ear–face biometric recognition approaches are proposed, one with fusion at the score-level and another at the feature-level. These approaches are based on local 3D features which are very fast to compute and robust to pose and scale variations, and occlusions due to hair and earrings. The fusion with ear data (which is not sensitive to changes in facial expression) significantly improves the face recognition results under non-neutral expressions. The complementary features of the two modalities also provide better results under neutral expression even without using the expensive ICP algorithm. The feature-level fusion approach performs better than the unimodal approaches based on ear or face only, but not better than the score-level fusion as intuitively believed. This is possibly due to the difficulty of fusing local ear and face features in a repeatable way. The performance of the proposed multimodal recognition can be improved further using a hybrid fusion approach where the result of a lower (e.g., data or feature) level of fusion can be fed to a higher (e.g., score) level of fusion. One possibility would be to short list gallery candidates using the result of data-level or feature-level fusion. A finer matching algorithm such as ICP can then be applied to each modality and at the end, score-level fusion can be used to make the final matching decision. Thus, we can combine the benefits of the different possible levels of fusion. Building a larger ear–face multimodal database with more challenging images (e.g., with eyeglasses) to evaluate the robustness of the proposed techniques can be another avenue of further research.

Acknowledgments

This research is sponsored by the Australian Research Council (ARC) under Grants DP0664228 and LED775672 and, by the University of Western Australia under Scholarships for International
Syed Islam received his BSc Electrical and Electronic Engineering and MSc Computer Engineering degrees in 2000 and 2006 respectively. He received a PhD with Distinction in Computer Engineering from The University of Western Australia (UWA) in 2011. He is currently a Research Assistant Professor at UWA.

Rowan Davies was born in Subiaco, Western Australia. He received a BCM (Hons) degree at the University of Western Australia in 1992, and a PhD degree at the Department of Computer Science, Carnegie Mellon University in 2005. Since 2001 he has been employed as a Lecturer at UWA.

Mohammed Bennamoun received his MSc from Queen’s University in the area of Control Theory, and his PhD from Queen’s /Q.U.T in the area of Computer Vision. He is currently a Professor in the School of Computer Science and Software Engineering at UWA. He was the Head of the School from February 2007 to April 2012.

Robyn Owens obtained a BSc (Hons) in Mathematics at the University of Western Australia (UWA) in 1974 before completing an MSc (1976) and a DPhil (1980), also in Mathematics, at Oxford University. She has lectured and researched in Mathematics and Computer Science. She is currently Deputy Vice-Chancellor (Research) at UWA.

Ajmal Mian received his BE Avionics and MS Information Security degrees in 1993 and 2002 respectively. He received PhD with distinction in Computer Science from UWA in 2006. He was awarded the Australian Postdoctoral Fellowship (2008) and Australian Research Fellowship (2011). He is currently a Research Associate Professor at UWA.