Review Article

An Overview of Multimodal Biometrics Using the Face and Ear

Yichao Ma , Zengxi Huang , Xiaoming Wang , and Kai Huang

School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

Correspondence should be addressed to Zengxi Huang; huangzx001@mail.xhu.edu.cn

Received 14 August 2020; Revised 17 September 2020; Accepted 4 October 2020; Published 31 October 2020

Academic Editor: Jianbing Ma

Copyright © 2020 Yichao Ma et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the recent years, we have witnessed the rapid development of face recognition, though it is still plagued by variations such as facial expressions, pose, and occlusion. In contrast to the face, the ear has a stable 3D structure and is nearly unaffected by aging and expression changes. Both the face and ear can be captured from a distance and in a nonintrusive manner, which makes them applicable to a wider range of application domains. Together with their physiological structure and location, the ear can readily serve as supplement to the face for biometric recognition. It has been a trend to combine the face and ear to develop nonintrusive multimodal recognition for improved accuracy, robustness, and security. However, when either the face or the ear suffers from data degeneration, if the fusion rule is fixed or with inferior flexibility, a multimodal system may perform worse than the unimodal system using only the modality with better quality sample. The biometric quality-based adaptive fusion is an avenue to address this issue. In this paper, we present an overview of the literature about multimodal biometrics using the face and ear. All the approaches are classified into categories according to their fusion levels. In the end, we pay particular attention to an adaptive multimodal identification system, which adopts a general biometric quality assessment (BQA) method and dynamically integrates the face and ear via sparse representation. Apart from a refinement of the BQA and fusion weights selection, we extend the experiments for a more thorough evaluation by using more datasets and more types of image degeneration.

1. Introduction

Face recognition (FR) has been intensively studied and received significant progress in the recent decade. Besides the better acceptability and the widely available low-cost camera sensors, compared with other popular biometrics such as fingerprint, iris, and retina recognition, face recognition has the potential to recognize uncooperative subjects from a distance and in a nonintrusive manner [1]. Therefore, face recognition can be applied to a wider range of applications, including biometric authentication, surveillance security, border control, forensics, and digital entertainment. However, owing to the variations such as facial expression, aging, pose, illumination, occlusion (e.g., sunglasses, scarf, and mask), and the increasing risk of spoof attacks, FR is not yet as accurate, flexible, and secure as desired [2, 3].

Recent studies have also validated that the ear can be used for biometric recognition. Compared with FR, the ear recognition (ER) has several appealing advantages: The ear

has a stable 3D structure with rich information, nearly unaffected by aging and facial expressions. ER is also contactless or nonintrusive. The ear is located near the face and can be captured along with the face using the same type of sensor or a single sensor at two times. Lots of popular face feature extraction and classification techniques are applicable to the ear. Therefore, it is highly attractive to combine the face and the ear and, thus, develop nonintrusive multimodal biometric systems for better recognition performance [2–5].

However, it must be noticed that data degeneration of either one of the modalities combined would degrade the multimodal recognition performance [2, 3]. Particularly, when one or a subset of the modalities used is corrupted severely to a certain extent, a multimodal system may perform worse than the unimodal systems using the other modality with a good quality sample. This consequence mainly results from the fact that most existing multimodal systems either use fixed fusion rules or their fusion rules cannot effectively adapt to the variations of biometric traits

preprocessing approaches, such as denoising and normalization, are generally employed before fusion. Chang et al. concatenate the normalized, masked ear, and face images of a subject to form a combined face-plus-ear image [11]. Jain and Ross [12] introduce a fusion scheme of mosaicking multiple impressions of the same finger to form an enhanced fingerprint. Sensor-level fusion could possibly make full use of all the evidence presented, but it is rarely used in the literature due to the fact that the raw biometric data may contain noisy or redundant data, together with its high computational complexity.

2.2. Feature-Level Fusion. Fusion at this level is performed after feature extraction but before matching. They consolidate the feature sets extracted from multiple biometric samples or by multiple algorithms into a single feature set. The common feature fusion techniques include serial concatenation [13], parallel fusion using a complex vector [14], and the methods that extract correlation feature of multiple modalities [15]. Compared with the latter two methods, serial concatenation is simple but effective and easy to extend for combining more than two modalities. Feature fusion schemes are expected to pertain most of the discriminative information from multiple biometric sources while containing less redundant data; thereby, they are expected to be the best way to improve multibiometric performance. However, the feature sets of multiple modalities may be incompatible; for example, the minutiae set of fingerprints and the eigen-coefficients of face are irreconcilable [8]. Besides, concatenating several feature vectors with fixed dimensionality might lead to the curse-of-dimensionality problem.

2.3. Score-Level Fusion. A match score represents the result of comparing two feature sets extracted using the same feature extractor. Each biometric matcher provides a match score. After transformed into a common domain, the match scores can be easily combined by using some simple rules such as Sum, Product, and Max or Min rules, or we concatenate all scores to form a score vector, and this vector is classified using Fisher's discriminant analysis, support vector machine (SVM), Bayesian classifier, neural network, and decision tree. Fusion approaches at this level are most commonly used in the biometric literature primarily due to the ease of accessing and processing match scores. In contrast to feature-level fusion, fusion at the score level is applicable to all kinds of multibiometric systems, while the information contained in matching scores is much richer than ranks and decisions. It is widely recognized that score-level fusion strikes the best balance between the effectiveness and the ease of fusion.

2.4. Rank-Level Fusion. Rank-level fusion is typically used in, but not limited to, multibiometric identification systems where each classifier associates a rank with every enrolled identity. Typically, a higher rank indicates a good match. The goal of rank-level fusion schemes is to consolidate the

ranking output by the individual biometric subsystems in order to derive a consensus rank for each identity. Ranks provide more insight into the decision-making process of the matcher compared to just the identity of the best match, but they reveal less information than match scores [8]. The ranks generated by different biometric matchers are comparable; thereby, rank-level fusion is simpler to implement than the score-level fusion, such as the highest rank method, the Borda count method, and the logistic regression method.

2.5. Decision-Level Fusion. Each biometric system makes its own recognition decision based on its own feature vectors or matching scores. While a verification system returns either a "match/accept" or a "nonmatch/reject" decision, an identification system makes such binary decision for every enrolled identity. In multibiometric systems, the binary decisions output by each matcher can be fused with AND or OR rules, majority voting, weighted majority voting, Bayesian decision fusion, and the Dempster-Shafer theory of theory. The least information contained in binary decision, compared with the other fusion approaches, implies that decision-level fusion is unlikely to achieve performance and popularity like score-level fusion.

Postclassification fusions are fairly popular due to the ease of accessing and processing the match scores, ranks, and decisions. Among them, score-level fusion is the most often seen in the literature. In contrast, combinations at the early stage are relatively difficult because the raw biometric data may contain noisy or redundant data, while feature sets extracted from different biometric traits may be incompatible. Nevertheless, because of the capability of utilizing more discriminative information for classification and the emerging advanced computational techniques, preclassification fusions have drawn much attention in the recent years. Feature-level fusion is generally believed to have potential on exploiting the most discriminative information contained in the raw data so as to pull the multibiometric performance to a higher level.

3. Face- and Ear-Based Multimodal Biometrics

According to literature [1–5], we summarize the benefits of the combination of the face and ear as follows: (a) The ear and the face are less correlated, and the combination of their discriminative information generally improves the recognition performance; (b) the ear is not affected by facial expressions and aging, and when the face is unreliable or even unavailable, the multimodal system can rely on the ear, and vice versa; (c) it is much more difficult to spoof both the face and the ear simultaneously than to spoof only one; (d) both the data collections are contactless, and thus, the multimodal recognition can still operate from a distance and in a nonintrusive manner; (e) they can share the same feature extraction and matching algorithms, whereby the derived feature sets are compatible for biometric fusion (f) benefiting from their physiological location and structure, the multimodal recognition enjoys a wider range of recognition angle in the yaw direction; (g) the ear and can be

captured along with the face by using the same type of sensor or a single sensor at two times; and (h) face detection helps speed up ear detection by offering an ear region of interest. Inspired by these attractive features, the face- and ear-based multimodal biometrics has attracted much attention in computer vision community. We classify the approaches in the literature according to the fusion levels and their operation modes, as shown in Table 1.

3.1. Sensor-Level Fusion Methods. Chang et al. [11] study the comparison and combination of ear and face biometrics with principal component analysis- (PCA-) based feature extraction algorithm. The probe images differ from the gallery in one of three categories: day variation (88 subjects), lighting variation (111 subjects), and pose variation (101 subjects). Every subject has two images: one is used as gallery image, and the other was used as probe image. Fusion is performed at the sensor level, where the normalized face and ear images of a subject were concatenated to form a combined face-plus-ear image. In the day variation experiment, multimodal method achieved 90% rank-one recognition rate, while the face's and the ear's are comparable and they are 70.5% and 71.6%, respectively. A significant progress is also achieved in lighting variation experiment. Face recognition in this experiment gets 64.9%, ear recognition gets 68.5%, and the multimodal recognition achieves 87.4%. Overall, the multimodal recognition gets a significant improvement of roughly 20% rank-one recognition rate over the FR and ER in both lighting variation and day variation experiments.

Yuan et al. [16] adopt the similar fusion scheme but use full-space linear discriminant analysis (FSLDA) instead of PCA for feature extraction. In experiments, the 4 ear images of a subject from the USTB ear database are randomly coupled with 4 face images of a subject from the ORL face database to form 4 multimodal biometric samples, of which 3 samples are used for gallery and the rest is used for probe. There are 75 subjects in total. As their experimental results show, face recognition gets 88.0% rank-one recognition rate, lower than the 94.7% of ear recognition. The accuracy is 98.7% when it comes to the multimodal recognition, which outperforms the biometric recognition using either one alone.

Yamen et al. [17] explore three approaches to fuse the profile and ear, namely, spatial fusion, intensity fusion, and channel fusion. In spatial fusion, they concatenate profile face and ear images side-by-side. In channel fusion, they stack the color channels of face and ear images; for example, the combined images will have 6 color channels if the input data is RGB images. In the intensity fusion, they average pixel intensity values of the profile face and ear images. After sensor-level fusion, they employ a fine-tuned CNN model, VGG-16 or ResNet-50, to extract the multimodal feature and classify the user age and gender simultaneously. In their experiments, compared with the state-of-the-art methods, the proposed method using channel fusion and VGG-16 achieves the best results on both age and gender classification tasks, which are, respectively, 67.59% and 99.11% on

two multimodal datasets selected from publicly available face datasets.

3.2. Feature-Level Fusion Methods. Xu and Mu [18] report a multimodal biometric system integrating the ear and profile face at the feature level. After PCA-based feature extraction, they gain the fused feature by means of canonical correlation analysis (CCA). A best accuracy of 97.37% is achieved on a subset of the USTB ear database with 38 subjects and 3 images/subject for gallery and 2 images/subject for probe. Furthermore, they develop a Kernel CCA-based feature fusion algorithm, which gets a further improved accuracy of 98.68% on the same database [19].

Abate et al. [5] employ iterated function systems transformation (IFS) to characterize the self-similarity of a face or ear image. Each face component, as well as the ear, is described by a list of centroids, which are, then, concatenated to form an overall feature vector. The system is tested on a subset with 100 subjects (2 images/subject) from the FERET database. When both the face and the ear probe images are occluded by blocks of 50×50 pixels, the multimodal system gets a 100% correct recognition rate; in contrast, the recognition rates of FR and ER based on IFS are only 90% and 81%, respectively. Their respective recognition rates are 93%, 80%, and 70% when the area of occluded blocks comes up to 100×100 pixels.

Theoharis et al. [20] propose a unified approach that fuses 3D face and ear data. They construct, in advance, an annotated face model (AFM) and an annotated ear model (AEM) based on statistical data. As a preprocessing step, an AFM or AEM is fitted to the face or ear 3D data by using ICP and Simulated Annealing (SA). Wavelet coefficients, as feature, are calculated from the resulted geometry images. For the multimodal fusion, the feature vector of each modality is multiplied by a global normalization weight, and then, all of them are concatenated into a single vector. The weights are selected empirically and slightly in favor of the face feature, as they consider the face modality more reliable. On a virtual multimodal database with 324 subjects based on the FRGC v2 face database (4007 range images, 466 subjects) and UND ear database (830 range images, 415 subjects), this multimodal method achieves 99.7% rank-one recognition rate, which is obviously better than the 97.5% for FR and 95% for ER.

Huang et al. [2] introduce a multimodal method called MSRC, which combines the PCA feature sets of the face and the ear with serial concatenation and employs an SR-based classification for multimodal classification. MSRC is reported with a significant improvement compared to Xu's CCA-based methods. However, they argue that although all existing multimodal systems are reported with evidently better performance than the unimodal recognition using either the face or ear alone, they may perform much worse than unimodal systems when the face or ear encounters image degeneration, owing to their fixed fusion rules. Their experiments show that even using advanced classification techniques, MSRC cannot avert a rapid decline of accuracy in the cases of face or ear data degeneration. In order to

TABLE 1: Face- and ear-based multimodal biometric methods in the literature.

Categories	Identification	Verification
Sensor-level fusion	Chang et al. [11] Yuan et al. [16] Yaman et al. [17] Abate et al. [5]	—
Feature-level fusion	Xu and Mu [18, 19] Theoharis et al. [20] Huang et al. [2–4] Yaman et al. [17]	Badrinath and Gupta [21]
Score-level fusion	Huang et al. [4] Xu and Mu [22] Yan [23] Mahoor et al. [24] Cadavid et al. [25] Yaman et al. [17]	Islam et al. [26] Mahoor et al. [24] Darwish et al. [27] Kisku et al. [28] Raghavendra et al. [29] Yazdanpanah et al. [30] Huang et al. [31] Huang et al. [32]
Rank-level fusion	—	Monwar and Gavrilova [33–35] Kumar et al. [36] Xu and Mu [22]
Decision-level fusion	Rahman and Ishikawa [37]	Kisku et al. [38] Boodoo and Subramanian [39]

handle this limitation, they propose a quality-based adaptive feature fusion scheme, thereby developing a multimodal approach called MRSCW [2]. Experiments demonstrate that MRSCW achieves quite encouraging robustness against image degeneration and outperforms many up-to-date methods. Very impressively, even when a query sample of one modality is extremely degenerated, MRSCW can still get a performance comparable to the unimodal recognition using the other modality.

In [17], Yamen et al. also explore the feature fusion strategy to combine the profile face and ear for age and gender classification. They first train two separate CNN models using the profile face and ear images. The CNN-based feature vectors of the two traits are concatenated to form a multimodal feature vector, which are, then, fed to the fully connected layers for simultaneous age and gender classifications. Although the network models are more sophisticated, both the feature fusion methods based on VGG-16 and ResNet-50 are inferior to the methods using channel fusion at the sensor level. The insufficient training samples used may be responsible for this result. Meanwhile, it is noted that the channel fusion methods preserve all the face and ear information to train the multimodal CNN models, while feature level fusion methods may have lost discrimination information before multimodal combination.

3.3. Score-Level Fusion Methods. Xu and Mu [22] combine the ear and face profile and use FSLDA feature extraction algorithm for both the ear and face. They test three score fusion schemes, i.e., Sum, Product, and Median rules, together with a decision-level fusion scheme. A subset of the USTB II ear database is used, which consists of 294 images for 42 subjects. Each subject has seven profile views head

images with variations of the head position and slightly facial expressions. In the experiments, for each subject, 5 images are used for gallery and the other 2 images are used as the probe image. As a result, 94.05% subjects can be recognized by ER and 88.10% subjects can be recognized correctly by FR. For the fusion schemes, the best performance is achieved by the Sum rule, while the Median and Product rules get accuracies of 97.62% and 96.43%, respectively.

Yan [23] explore multimodal biometrics by using face and ear 3D data. Experiments are performed on a database with 174 subjects, and each has two ear shapes and two face shapes. The proposed system uses an improved ICP-based (Iterative Closest Point) approach and is fully automatic. The recognition rates of FR and ER are 93.1% and 97.7%, respectively, while the multimodal recognition with Sum rule yields 100% recognition rate.

tMahoor et al. [24] combine the 2.5D ear data and 2D face image at the score level. For 2.5D ear recognition, a series of frames is extracted from a video clip. The ear segment in each frame is independently reconstructed using the shape from the shading method. Then, various ear contours are extracted and registered using the ICP algorithm. For 2D face recognition, a set of facial landmarks are extracted from frontal facial images using active shape model. Then, the responses of facial images to a series of Gabor filters at the locations of facial landmarks were calculated and are used for recognition. They report accuracies of 81.67%, 95%, and 100% for face, ear, and fusion, respectively, on the WVU database.

Islam et al. [26] fuse 3D local features for ear and face at the score level, using weighted sum rules. They use the FRGC v2 3D face database and UND ear databases. The proposed multimodal method achieves an identification rate of 98.71% and a verification rate of 99.68% (at 0.001 FAR) for neutral

face expression. For other types of facial expressions, they achieve 98.1% and 96.83% identification and verification rates, respectively.

Huang et al. [31] propose a face- and ear-based multimodal verification system by using sparsity-based matching metric. They construct dynamic dictionary by using the training samples of the claimed client and some nontarget subjects. The face and ear query samples are first encoded separately, and then, the resulting sparsity-based matching scores are combined with Sum-rule fusion for multimodal verification. They consider the sparse coding as a competing one-to-many matching process where a verification request can be accepted only in the case when the genuine class defeats almost all the nontarget classes and achieves an eligible sparsity-based matching score in encoding the query data. Hence, the system not only examines the matching score but also compares implicitly the correlations of the query data to the client and many nontarget subjects and, thereby, offers double insurance for identity security. Their experiments demonstrate that the proposed multimodal method is not only better than its unimodal counterparts but also significantly superior to the well-known multimodal methods such as likelihood ratio (LLR), support vector machine (SVM), and the sum-rule fusion methods using cosine similarity.

Furthermore, to enhance the resistance against the face or ear unimodal spoof attacks, Huang et al. [32] propose to use the collaborative representation fidelity to measure the anomaly degree of a query sample to the claimed client. They combine the anomaly degrees and sparsity-based matching scores obtained from the face and the ear query samples in a stack way. In the end, they use the genuine, imposter, and partial spoofed multimodal score samples to train a SVM classifier for multimodal verification. Extensive experimental results demonstrate the superiority of the proposed method in licit scenario and under the worst-case partial spoof attacks. More importantly, the proposed method achieves a good balance between the accuracy and security when the system uses a fixed operating threshold for both licit and spoofing scenarios.

In [17], Yamen et al. also evaluate the score level fusion for multimodal age and gender classification using profile face and ear. They use two individual CNN models to generate the probabilities associated with all age or gender classes. Then, they achieve the confidence scores by using a certain calculation method with the the probabilities as input data. They finally select the prediction of the CNN model that has the maximum confidence score. In their experiments, they evaluate the proposed score fusion approach with a series of confidence calculation methods, but both the age and gender classification results are inferior to the proposed sensor- and feature-level fusion methods.

3.4. Rank-Level Fusion Methods. Monwar and Gavrilova [33] combine the face, ear, and signature for identity verification by utilizing rank-level fusion approaches, i.e., the highest rank, Borda count, and logistic regression. As to feature extraction, the same PCA or Fisher's linear

discriminant (FLD) approach is employed for all modalities. They build a chimeric database consisting of faces, ears, and signatures for testing. For the face, they use the ORL database, containing 400 images, 10 each of 40 different subjects. For the ear, they use the Carreira-Perpinan database. For the signature, they use 160 signatures of the Rajshahi database. The experimental results indicate that fusion with weighted Borda count can improve the overall equal error rate (ERR) to 9.76% compared to an average of 16.78% for verification with individual modalities. Later, Monwar and Gavrilova [34] extend their experiments by including more data from the USTB database. For the signatures, they use 500 signatures in total, with 10 signatures/subject, from the Rajshahi database. They achieve an EER of 1.12% when using the logistic regression rank-level fusion scheme.

3.5. Decision-Level Fusion Methods. Rahman and Ishikawa [37] put forward a multimodal biometric verification system using ear and face profile. For each modality, they utilize PCA to extract the features and do classification individually. If the multimodal system recognizes any one of the ear and face of a particular person successfully, they consider the correct identification of the subject. This is a typical decision level fusion scheme called OR rule. In their experiments on a database with 18 subjects and 5 images per subject, they achieve a best recognition rate of 94.44% (by adjusting the threshold value), while the best of face and ear unimodal verification methods get 88.88% and 77.77%, respectively.

Xu and Mu [22] propose a decision-level fusion scheme called the Modified-Vote rule to combine the decision results of ear and face profile subsystems. The modified-Vote rule is reported slightly inferior to score-level fusion schemes such as Sum and Product rules, while it is comparable to the Median-rule. Kisku et al. [38] apply two trained Gaussian Mixture Models (GMM) to estimate the match score distributions of Gabor features of the face and ear, respectively, and, then, verify an identity based on the Dempster-Shafer theory of evidence. They get an EER of 4.47% on the IIT Kanpur multimodal database, having 400 subjects in total.

4. Adaptive Face- and Ear-Based Multimodal Fusion

Quality-based adaptive multimodal fusion is an intuitive but very effective way to improve the multimodal recognition performance in the cases when one or a subset of modalities suffers from data generation. In [2], Huang et al. propose a general biometric quality assessment method for the face and the ear by means of sparse representation. By taking advantages of BQA and sparse representation-based classification, they develop an adaptive multimodal identification system that is able to dynamically integrate the face and ear features. When the face or ear image degeneration is severe to a certain extent, the multimodal system can effectively reduce its negative effect. Their experiments demonstrate that when the face (ear) query sample suffers from 100% random pixel corruption, the multimodal system

can still get the performance close to the ear (face) unimodal recognition. Moreover, the employment of BQA and the related dynamic feature fusion does not lead to efficiency reduction because the biometric quality is derived from the coding results of the adopted sparse representation-based classification. In this section, we will introduce the adaptive multimodal system in details and refine its BQA and fusion weights selection partly according to literature [3]. We also evaluate the BQA and the overall performance against more types of image degeneration, including random pixel corruption, random block occlusion, real face disguise (i.e., sunglasses and scarf), and illumination variation.

4.1. Sparse Representation-Based Classification. Suppose that there are c classes of subjects and m samples per class, and let $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_c]$ ($\mathbf{A} \in \mathbb{R}^{N \times mc}$, $N < mc$) be the over-complete dictionary, where $\mathbf{A}_i = [\mathbf{a}_{i,1}, \mathbf{a}_{i,2}, \dots, \mathbf{a}_{i,m}]$ ($i = 1, 2, \dots, c$) is a subset of training samples of class i . Let $\mathbf{y}$ be a query sample. The original SRC algorithm is summarized as follows [40, 41].

- (a) We encode $\mathbf{y}$ over $\mathbf{A}$ via L_1 -norm minimization

$$\begin{aligned} \hat{\mathbf{a}} &= \arg \min \|\mathbf{a}\|_1, \\ \text{s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{a}\|_2 &< \varepsilon, \end{aligned} \quad (1)$$

where $\varepsilon > 0$ is a constant.

- (b) We compute the class-specific sparse representation error for each class

$$\eta_i = \|\mathbf{y} - \mathbf{A}_i \hat{\mathbf{a}}_i\|_2, \quad (2)$$

- (i) where $\hat{\mathbf{a}}_i$ is the coefficients associated with class i .

- (c) We perform classification via $O(\mathbf{y}) = \arg \min_i \{\eta_i\}$.

Yang et al. [42] argue that the original SRC model is based on the assumption that the coding residual follows Gaussian or Laplacian distribution, which may not hold well in practice, especially when occlusion and corruption occur in the query image. Hence, they develop a new model, namely, the robust sparse coding model (RSC), to seek for the maximum likelihood estimation (MLE) solution to the sparse coding problem. Compared with the SRC model, RSC can estimate and suppress the outliers (e.g., the image pixels or regions corrupted or occluded) by dynamically assigning them with lower weights. The RSC model can be defined by

$$\begin{aligned} \min_{\mathbf{a}} \quad & \|\mathbf{W}_{\text{pw}}^{1/2} (\mathbf{y} - \mathbf{A}\mathbf{a})\|_2^2, \\ \text{s.t. } \quad & \|\mathbf{a}\|_1 \leq \sigma, \end{aligned} \quad (3)$$

where $\sigma > 0$ is a constant and $\mathbf{W}_{\text{pw}}^{1/2} = \text{diag}(\omega_1, \omega_2, \dots, \omega_N)$ ($\omega_k \in [0, 1]$) is a diagonal

matrix with nonnegative scalars. The subscript “pw” represents “pixel weight”.

The reconstruction residual is $\mathbf{r} = \mathbf{y} - \mathbf{A}\hat{\mathbf{a}} = [r_1, r_2, \dots, r_N]$. Let $\mathbf{r}' = [r_1^2, r_2^2, \dots, r_N^2]$, and we reorder its elements in an ascending order to be a new vector $\mathbf{r}'' = [r_1', r_2', \dots, r_N']$ ($r_k' \leq r_{k+1}'$). Then, ω_k is selected with a logistic function as

$$\omega_k = \frac{1}{1 + \exp(\mu((r_k'/\delta) - 1))}, \quad (4)$$

where μ is a positive constant that controls the decreasing rate and parameter δ controls the location of demarcation point, which is defined by $\delta = r''(\lfloor \tau N \rfloor)$, where $\tau \in (0, 1]$.

4.2. Biometric Quality Assessment. It is generally acknowledged that a biometric quality metric should be able to mirror the utility of a biometric sample, which refers to the impact of the individual biometric sample on the overall performance of a biometric system [43, 44]. However, how to assess and measure the utility of a biometric sample is often difficult because there are a variety of factors that may degrade the biometric quality in practice, such as illumination, pose, expression, and aging effect in face recognition. Assembling multiple quality measures, e.g., signal to noise ratio (SNR), mean squared error (MSE), image resolution, and sharpness, may come to a more comprehensive representation of biometric quality. Nevertheless, this may result in a complex quality-based fusion scheme or classifier and, hence, leads to a higher chance of overfitting. Thus, it is desirable to seek for a uniform biometric quality measure that is applicable to various degrading factors. Moreover, as for an adaptive multimodal system that intends to eliminate the adverse effect of modalities with a low-quality sample, the access to biometric quality difference among samples of the modalities is critical to its effectiveness. Therefore, the compatibility of quality assessment results of different modalities is very important.

The success of sparse representation is generally believed to benefit from a simple but important property of the natural data: although the high-dimensional images (or their features) belonging to the same class exhibit degenerate structure, they lie on or near low-dimensional subspaces, submanifolds, or stratifications [45]. Generally, a typical face/ear query image of a subject enrolled can be expected to be encoded with high fidelity over the dictionary $\mathbf{A}$ via L_1 -norm minimizations in equation (1) or equation (3). However, in many applications, the query samples are often degenerated by a variety of factors such that they may not lie on or near the targeted low-dimensional subspace spanned by the atoms of the established dictionary. In this context, it is not reasonable to expect a high fidelity of sparse representation. Nevertheless, it is still possible to seek a linear combination of dictionary atoms as close as possible to represent the query sample, using a relatively relaxed L_1 -norm sparsity constraint. As a result, the overall representation error $\rho = \|\mathbf{y} - \mathbf{A}\mathbf{a}\|_2$ would be rather evident.

Motivated by the abovementioned observations, Huang et al. [2] propose to use the overall representation error as a biometric quality measure for the face and the ear, namely, collaborative representation fidelity (CRF). Note that some corruptions such as isolated/impulsive noise and small occlusion regions often cause large coding residual but lead to a relatively small impact on biometric recognition. Therefore, aiming to more effectively predict biometric performance, they revise the CRF quality measure by using a certain part of coding residual.

Suppose that $\mathbf{y} \in R^N$ is a query sample and $\mathbf{A} \in R^{N \times mc}$ is an overcomplete dictionary consisting of all training samples. With the L_1 -norm optimization result of equation (1) or equation (3), the reconstruction residual is obtained by $\mathbf{r} = \mathbf{y} - \mathbf{A}\hat{\alpha} = [r_1, r_2, \dots, r_N]$. Let $\mathbf{r}' = [r_1^2, r_2^2, \dots, r_N^2]$, and its elements are reordered in an ascending order to be a new vector $\mathbf{r}'' = [r_1', r_2', \dots, r_N']$ ($r_k' \leq r_{k+1}'$). Then, CRF can be computed with the first n elements of $\mathbf{r}''$ as follows:

$$\rho = \sqrt{\sum_{i=1}^n r_i'} \quad (5)$$

Figures 2(a) and 2(c) demonstrate that, in both SRC-based FR on the AR face database [46] and ER on the USTB III ear database [47], all the CRF values are very small and around 0.04. On the other hand, when the probe face and ear images suffer from random pixel corruption, replacing a certain percentage of image pixels by uniformly distributed random values within $[0, 255]$, the CRF values consistently increase with the rise in the percentage of pixels corrupted. The details of experiment settings will be given in 4.5.

Table 2 provides the rank-one recognition rates of SRC-based FR and ER against various levels of random pixel corruption. Both accuracies of the face and ear decrease drastically with the enhancement of corruption. Specifically, the performance of ear biometric is more sensitive to the corruption, which is in accordance with the comparison between Figures 2(b) and 2(d) that the CRF of ear increases faster than that of the face. This detail strengthens the correlation between the biometric performance and the CRF value.

The abovementioned experimental results demonstrate that the CRF is able to mirror the utility of the face and ear samples that are degenerated by random pixel corruption or pose variation. In Section 4.5, more evidence to support the feasibility of CRF will be found in the experiments under conditions such as illumination changes, sunglasses and scarf occlusion, and random block occlusion.

4.3. Quality-Based Fusion Weights Selection. The basic idea of the quality-based adaptive multimodal fusion in [2] is to assign a punitive weight smaller than 1 to the less reliable modality according to the CRF quality. They assess the quality difference between the face and ear

query samples by their CRFs ratio, formulated in the following equation:

$$\begin{aligned} \gamma^{fe} &= b \cdot \frac{\rho^f}{\rho^e}, \\ \gamma^{ef} &= \frac{\rho^e}{b \cdot \rho^f}, \end{aligned} \quad (6)$$

where b is a balance factor, which is set as 1.0 if there are no specific instructions.

The γ^{fe} (γ^{ef}) represents the ratio of the face (ear) CRF to the ear (face) CRF. A γ^{fe} (γ^{ef}) larger than 1 means the lower quality of the face (ear) query sample relative to the ear's (face's). Generally, when γ^{fe} (γ^{ef}) is larger than a certain threshold θ ($\theta > 1$) and, meanwhile, $\rho^f > \rho_{\max}$ ($\rho^e > \rho_{\max}$, $\rho_{\max}$ is a threshold), the face (ear) feature or score should be assigned with a weight $w^f < 1$ ($w^e < 1$), and meanwhile, let $w^e = 1$ ($w^f = 1$). Ideally, the weight w^f (w^e) should be a monotone decreasing function with a variable of γ^{fe} (γ^{ef}). A uniform fusion weights selection function for the face and the ear, i.e., w^f (γ^{fe}) and w^e (γ^{ef}), is formulated with a logistic function as follows:

$$w(\gamma) = \begin{cases} 1, & 0 < \gamma \leq \theta, \\ \frac{2}{1 + \exp((\gamma^2 - \theta)/\varphi)}, & \theta < \gamma, \text{Max}(\rho^f, \rho^e) > \rho_{\max}. \end{cases} \quad (7)$$

4.4. System Architecture. The quality-based weighting scheme is designed to alleviate the adverse effect of the lower quality modality. It may not help if both the face and ear samples are degenerated at a similar level. Thus, to enhance the robustness to simultaneous data degeneration, Huang et al. adopt the RSC model for sparse coding, which helps suppress the outlier pixels or regions in the images by using an iterative coding scheme [42]. Their strategy is that the quality-based fusion scheme is utilized to handle the unimodal data degeneration cases, while the RSC model is adopted as a supplementary measure for tackling the multimodal data degeneration cases. They finally propose a quality-based adaptive multimodal RSC identification system, namely, AMRSC in short, whose block diagram is shown in Figure 3. AMRSC integrates the face and ear features using the serial concatenation and performs sparse coding at the feature level. It actually uses a two-level weighting strategy, including the pixel level and feature level. It seeks to suppress the outlier pixels at the former level and reduce the adverse effect of the less reliable modality at the latter level.

Suppose that there are c subjects and m samples per subject for both the face and ear. Let $\mathbf{A}^f = [\mathbf{A}_1^f, \mathbf{A}_2^f, \dots, \mathbf{A}_c^f] \in R^{N \times mc}$ and $\mathbf{A}^e = [\mathbf{A}_1^e, \mathbf{A}_2^e, \dots, \mathbf{A}_c^e] \in R^{N \times mc}$ ($N < mc$) separately denote training datasets of

FIGURE 2: CRF distributions derived from SRC-based FR and ER. (a) and (c) are under common conditions; (b) and (d) are under different levels of random pixel corruption.

TABLE 2: SRC-based unimodal recognition accuracy (%) versus random pixel corruption.

Corruption	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
AR	94.394	92.586	86.799	76.492	60.94	46.293	23.327	12.116	5.606	3.074	1.266
USTB III	88.023	80.526	69.815	54.236	37.585	22.493	14.606	7.011	3.505	2.434	1.168

FIGURE 3: Block diagram of the adaptive multimodal identification system.

the face and ear, and $\mathbf{A} = [\mathbf{A}^f; \mathbf{A}^e]$ is the multimodal overcomplete dictionary. Let $\mathbf{y}^f$ and $\mathbf{y}^e$ be the face and ear query samples, and $\mathbf{y} = [\mathbf{y}^f; \mathbf{y}^e]$ is the multimodal query data. If the PCA projection matrices of the face and ear are $\mathbf{P}^f$ and $\mathbf{P}^e$ and $\mathbf{P} = [\mathbf{P}^f, \mathbf{P}^e]$, then the features of training samples can be obtained by $\mathbf{D}^f = (\mathbf{P}^f)^T \mathbf{A}^f$ and $\mathbf{D}^e = (\mathbf{P}^e)^T \mathbf{A}^e$, and thereby, their multimodal fusion is $\mathbf{D} = [\mathbf{D}^f; \mathbf{D}^e]$. In the same way, the features of query samples and their fusion are $\mathbf{z}^f = (\mathbf{P}^f)^T \mathbf{y}^f$, $\mathbf{z}^e = (\mathbf{P}^e)^T \mathbf{y}^e$, and $\mathbf{z} = [\mathbf{z}^f; \mathbf{z}^e]$, respectively.

For simplicity, the quality-based fusion weights for the face and ear are formulated as follows:

$$\mathbf{W}_{fw} = \begin{bmatrix} \mathbf{W}_{fw}^f & 0 \\ 0 & \mathbf{W}_{fw}^e \end{bmatrix}, \quad (8)$$

where $\mathbf{W}_{fw}^e = \text{diag}(w^e, w^e, \dots, w^e)$ and they have the same dimensions with $\mathbf{z}^f$ and $\mathbf{z}^e$, respectively.

Finally, the multimodal sparse coding problem in AMRSC can be formulated as

$$\begin{aligned} \min & \|\mathbf{W}_{fw} \mathbf{P}^T \mathbf{W}_{pw}^{1/2} (\mathbf{y} - \mathbf{A}\boldsymbol{\alpha})\|_2^2 \\ \text{s.t. } & \|\boldsymbol{\alpha}\|_1 \leq \sigma. \end{aligned} \quad (9)$$

Equation (9) can be simplified to be

$$\begin{aligned} \min & \|\mathbf{W}_{fw} (\mathbf{z} - \mathbf{D}\boldsymbol{\alpha})\|_2^2 \\ \text{s.t. } & \|\boldsymbol{\alpha}\|_1 \leq \sigma, \end{aligned} \quad (10)$$

where $\mathbf{z} = \mathbf{P}^T \mathbf{W}_{pw}^{1/2} \mathbf{y}$, $\mathbf{D} = \mathbf{P}^T \mathbf{W}_{pw}^{1/2} \mathbf{A}$.

The L_1 -norm optimization problem in equation (10) can be solved by $l_1 - l_s$ [48]. AMRSC performs sparse coding in multimodal feature space. The resulting code vector $\hat{\boldsymbol{\alpha}}$ is utilized to estimate the pixel weight matrix $\mathbf{W}_{pw}$ with equation (4), as well as the CRF quality. Similar with the algorithm in solving the RSC model, AMRSC uses an iterative sparse coding process. For each iteration, the outliers of face and ear query images are gradually detected and suppressed, and the biometric qualities are accessed. The face and ear features are integrated dynamically since the second iteration.

4.5. Experiments and Discussions

4.5.1. Databases and Settings. The multiview USTB III ear database [47] and face databases including the Extended Yale B [49], Georgia Tech (GT) [50], and AR (the first 79 subjects) [46] are adopted to build three multimodal databases, namely, MD I, MD II, and MD III for short. Sample face and ear images are shown in Figure 4. Table 3 provides the compositions of multimodal databases. MD I and MD II use the first 38 and 50 subjects of USTB III. For each virtual subject, we pair 7 face images with 7 ear images to form 7 multimodal samples for training. Each multimodal training sample is a unique pair of face and ear samples from the same virtual subject. On the contrary, for obtaining more instances for testing, each face probe image is paired with all ear probe images, for a subject. For example, in MD I each subject has 38 face images and 13 ear images for testing, so

$C_{38}^1 C_{13}^1$ virtual multimodal instances can be obtained. In the experiments, all images of both modalities are normalized to 50×40 pixels. The PCA-based feature dimensionalities for both modalities on MD I, II, and III are 100, empirically selected as 150 and 200, respectively.

AMRSC is compared with seven multimodal methods, including MRSC, MSRCef, MSRCs, MCCA (Multimodal CCA) [18], MSVM (Multimodal SVM, using a polynomial kernel $(k(x, x') = 0.05 \times \langle x, x' \rangle + 1)^3$), MNFL (Multimodal Nearest Feature Line), and MNN (Multimodal Nearest Neighbor). MRSC is a special model of AMRSC when feature weight matrix $\mathbf{W}_{fw}$ is a unit matrix, or say, without quality-based feature weighting scheme. MSRCef is also a special model of AMRSC when both the weight matrices $\mathbf{W}_{pw}$ and $\mathbf{W}_{fw}$ are unit matrices. MSRCs combine the class-specific sparse representation errors of the face and ear with Sum-rule fusion for classification. MSVM, MNFL, and MNN combine the face and ear features with serial concatenation before classification. All the L_1 -norm optimization problems in AMRSC, MSRCef, and MSRCs are solved by using $l_1 - l_s$ [48]. The iteration number of AMRSC and MRSC are both two, while the parameter τ is set as 0.85 for them. Let $b = 1.0$, $\theta = 1.2$, $\rho_{\max} = 0.02$, and $\varphi = 3$ for the quality-based fusion weights selection of AMRSC.

4.5.2. Experiments. We conduct a series of experiments including six categories, i.e., controlled environment, illumination variation, face disguise (sunglasses and scarf), random pixel corruption, random block occlusion, and multimodal degeneration. For all the methods, the experiments under conditions of random pixel corruption and random block occlusion are repeated five times and the average accuracies are used as the final results. The recognition accuracy (or rate) hereafter is referred to the rank-one recognition rate.

- (i) Controlled environment: the face and ear unimodal recognition methods are used as baseline. On MD I, probe subset 4 is not used because of its severe degeneration, which will be used in the illumination test. Correspondingly, its face part, i.e., subset 5 of Yale B, does not join in the face recognition. As shown in Table 4, the SR-based methods, RSC and SRC, significantly outperform the other methods in both FR and ER. RSC performs the best on all datasets due to its outlier detection ability.

Table 5 summarizes the multimodal recognition results on the three multimodal databases. Compared with their unimodal counterparts, all the SR-based methods significantly improve the recognition performance. Meanwhile, their superiority to other multimodal competitors is significant. For instance, compared with MNN, all of them are able to get an accuracy increase over 20%. AMRSC achieves the best results on MD I and MD II. Although the advantage of AMRSC is not evident, it is reasonable that quality-based fusion is not designed for the cases where the face or ear query image is not obviously degenerated.

- (ii) Illumination variation: MD I is used for testing the robustness of multimodal methods against

FIGURE 4: Sample images of one person in the face and ear databases. (a) Yale B Subset 1, 2, 3, 4, 5. (b) Georgia Tech. (c) AR (Session 1, Session 2). (d) USTB III.

TABLE 3: The composition of MD I, II, and III.

Multimodal database (subjects)	Instance	Face database	Ear database
MD I (38)	Gallery	$7 \times 38 = 266$	Gallery (38)
	Subset 1	$12 \times 13 \times 38 = 5928$	
	Subset 2	$12 \times 13 \times 38 = 5928$	
	Subset 3	$14 \times 13 \times 38 = 6916$	
	Subset 4	$19 \times 13 \times 38 = 9386$	
MD II (50)	Gallery	$7 \times 50 = 350$	Gallery (50)
	Probe	$8 \times 13 \times 50 = 5200$	Probe (50)
MD III (79)	Gallery	$7 \times 79 = 553$	Gallery (79)
	Subset 1	$7 \times 13 \times 79 = 7189$	
	Subset 2	$6 \times 13 \times 79 = 6162$	
	Subset 3	$6 \times 13 \times 79 = 6162$	
	Subset 4	$6 \times 13 \times 79 = 6162$	

TABLE 4: Unimodal recognition accuracy (%) on MD I, II, and III.

	Face recognition			Ear recognition		
	Yale B	GT	AR	38 subjects	50 subjects	79 subjects
NN	44.598	74.5	73.599	81.846	80.721	77.733
NFL	48.753	81.25	75.769	82.794	82.794	78.745
SVM	46.607	82.25	74.141	84.008	84.008	80.972
SRC	91.066	90.5	94.394	89.271	88.462	88.057
RSC	91.136	93.25	95.841	92.713	92.308	92.105

illumination variation in face images. As the face part of MD I, the images of Yale B are more and more challenging from Subset 1 to Subset 5, as shown in Figure 4. As a result, from the Probe

Subset 1 to Probe Subset 4, the multimodal data quality becomes more and more inferior. Table 6 collects the multimodal recognition results. All the methods perform well when the face images are of good quality, such as on the Probe Subset 1 and 2. However, on Probe Subset 3 and 4, MSRCef, MSRCs, and MRSC get evident performance degradation. They are all obviously inferior to their counterparts in ER; that is, their accuracies are lower than the 89.271% of SRC-based ER (refer to Table 4). Similarly, in spite of the outlier suppression scheme, MRSC is much inferior to the RSC with the ear. On the contrary, with the help of quality-based feature fusion, AMRSC gets quite encouraging results, which

TABLE 5: Multimodal recognition accuracy (%) in controlled environment.

	MNN	MNFL	MSVM	MCCA	MSRCef	MSRCs	MRSC	AMRSC
MD I	72.059	75.293	77.221	90.758	92.862	97.001	93.826	96.58
MD II	85.577	88.981	90.404	85.558	97.365	97.75	98.019	98.135
MD III	75.129	78.133	87.717	91.946	98.345	98.873	98.512	98.998

TABLE 6: Multimodal recognition accuracy (%) versus illumination variation on MD I.

	MNN	MNFL	MSVM	MCCA	MSRCef	MSRCs	MRSC	AMRSC
Subset 1	94.585	96.592	98.043	98.026	100	99.781	100	100
Subset 2	77.109	81.545	84.733	96.778	99.882	99.291	98.212	97.807
Subset 3	48.424	51.677	52.935	79.367	80.726	92.655	84.774	93.276
Subset 4	36.533	38.568	36.981	53.569	23.876	76.603	66.866	92.531

FIGURE 5: Sample face and ear images with random pixel corruption.

TABLE 7: Multimodal recognition accuracy (%) versus real face disguise.

	MNN	MNFL	MSVM	MCCA	MSRCef	MSRCs	MRSC	AMRSC
Subset 2	67.04	72.866	82.554	27.004	86.482	93.411	96.414	96.186
Subset 3	36.254	41.285	59.964	48.442	87.796	96.284	91.951	95.862

are 93.276% and 92.531% on the last two subsets, respectively, compared with the 92.105% of ER with RSC. This result shows that AMRSC is able to reduce the adverse effect of severe illumination changes on the face images.

- (iii) Real face disguise: on Subset 2 and Subset 3 of MD III, the faces are disguised by sunglasses or a scarf. As shown in Figure 5, sunglasses almost cover 25% area of the face image, while the coverage of a scarf is about 40%. On Subset 2, RSC can effectively detect and suppress the sunglasses region because of its evident contrary with the same region in the common face image. As shown in Table 7, MRSC achieves the best accuracy of 96.414%. AMRSC also benefits from the RSC model and gets a comparable accuracy of 96.186%. The slight disadvantage of AMRSC over MRSC could probably be explained by the fact that, apart from suppressing the sunglasses region, AMRSC would also reduce the importance

of face modality based on its poor quality, while MRSC could make full use of the remainder information of the face image.

On Subset 3, it is more challenging for RSC to detect the scarf region on the face. MRSC only gets an accuracy of 91.951%, while the ER with RSC gets 92.105%. On the other hand, with the advantage of quality-based adaptive fusion, AMRSC gets an evident improvement of 3.76% compared with the ER with RSC. Let us say AMRSC can still take advantage of the discriminative information of the disguised face image. Although AMRSC fails to get the best results on both datasets, on the whole, it is more stable than the competitors.

- (iv) Random pixel corruption: by applying random pixel corruption to the face or ear images of Probe Subset 1 of MD III and MD II, we can evaluate the multimodal methods at different corruption levels, i.e., 20%, 40%, 60%, 80%, and 100%. Figure 5 shows sample images with these corruption levels. We call

TABLE 8: Multimodal recognition accuracy (%) versus random pixel corruption on MD III.

Corruption	0%	20%	40%	60%	80%	100%
<i>(a) Face corruption</i>						
MNN	75.129	71.595	67.242	59.021	47.531	30.227
MNFL	78.133	72.374	67.909	63.194	47.114	30.964
MSVM	87.717	87.091	85.13	78.717	64.946	45.194
MCCA	91.946	78.008	51.774	26.151	12.102	6.232
MSRCef	98.345	97.093	95.006	89.22	77.813	59.452
MSRCs	98.873	97.719	95.187	90.75	84.226	75.031
MRSC	98.512	97.148	95.41	91.084	90.291	87.69
AMRSC	98.998	98.317	95.284	93.337	92.475	92.141
<i>(b) Ear corruption</i>						
MNN	75.129	69.245	47.976	27.862	19.224	15.579
MNFL	78.133	72.138	48.004	28.391	20.573	17.012
MSVM	87.717	75.796	31.785	16.247	11.156	8.624
MCCA	91.946	83.642	73.306	63.722	54.013	48.157
MSRCef	98.345	97.454	94.728	89.901	76.603	63.903
MSRCs	98.873	98.025	94.505	89.025	80.581	70.135
MRSC	98.512	97.691	96.634	90.958	88.079	89.428
AMRSC	98.998	98.859	97.26	96.467	96.133	95.549

FIGURE 6: Multimodal recognition accuracy (%) versus random pixel corruption on MD II. (a) Face corruption. (b) Ear corruption.

FIGURE 7: Sample face and ear images with random block occlusion.

FIGURE 8: Multimodal identification accuracy (%) versus random block occlusion on MD III. (a) Face occlusion. (b) Ear occlusion.

FIGURE 9: Multimodal recognition accuracy (%) versus simultaneous random pixel corruption. (a) MD II; (b) MD III.

the cases when only the face or the ear image suffers from corruption/occlusion as unimodal corruption/occlusion.

Table 8 summarizes all the multimodal recognition results on MD III. In experiments of both types of unimodal corruption, inevitably all multimodal methods' performance decreases more or less with the corruption in increase. However, in the face corruption case, AMRSC achieves the

best accuracy at all corruption levels except the 40% corruption. Even at 100% face corruption, compared to the ER using RSC with 92.105%, AMRSC can still obtain a comparable accuracy of 92.141%. With the help of outlier detection, MRSC is much better than MSRCef and MSRCs, but it is evidently inferior to AMRSC. The advantage of AMRSC over MRSC clearly tells the effectiveness of quality-based fusion. AMRSC's superiority to other methods is overwhelming in the ear corruption case. When the ear

FIGURE 10: Multimodal recognition accuracy (%) versus simultaneous random block occlusion. (a) MD II; (b) MD III.

corruption is 100%, AMRSC's accuracy is 95.549%, which is very close to the 95.841% of FR with RSC. Overall, AMRSC is validated to have capability of tolerating the 100% image corruption of face or ear, and on the other hand, the effectiveness of quality-based fusion is proved again.

On MD II, we get similar results with those obtained on MD III. Figure 6 shows the recognition accuracy curves of all multimodal methods in unimodal corruption experiments. A dashed line representing the RSC-based unimodal recognition with the modality without corruption is used as a baseline for evaluating AMRSC's effectiveness. We can see that all the curves obtained by the multimodal methods with the traditional classification descend dramatically with the increase in the corruption level. MSRCef and MSRCs are able to tolerate about 50% corruption. In contrast, the AMRSC's curve is very stable, descending much more gently. AMRSC's curve even does not ever break the dashed line in the face corruption case, while although this happens in the ear corruption case, it has never been far away from the dashed line. These experiments on MD II demonstrate again that AMRSC can tolerate the most extreme random pixel corruption to the face or the ear image.

- (v) Random block occlusion: we simulate the random block occlusion by using a Baboon picture with various sizes, as shown in Figure 7. Figure 8 shows their performance under the conditions when only the face or the ear is occluded on MD III. While all other methods suffer from significant performance reduction with the aggravating occlusion, AMRSC can still achieve performance comparable to the ER with RSC when the 100% area of the face or the ear image is occluded. The competitors seem more sensitive to the ear random block occlusion. The performance of

MRSC in this series of experiments reveals that the RSC is not quite effective in dealing with complex occlusion. This, in turn, illustrates the importance of quality-based adaptive fusion in AMRSC.

- (vi) Multimodal degeneration: Figures 9 and 10 show the multimodal performance in the cases when the face and the ear simultaneously suffer from random pixel corruption and random block occlusion. It can be observed that all the accuracy curves of the multimodal methods with the traditional classification descend sharply after 20% corruption in both degeneration scenarios. AMRSC and MRSC are always comparable. This tells that the quality-based adaptive fusion does not help improve the performance when the face and ear encounter the same level of data degeneration. But, at least, it does not lead to an adverse effect. On the other hand, compared with the multimodal methods with the SRC model, their superiority is evident. They are able to tolerate 40% simultaneous random pixel corruption and 20% simultaneous random block occlusion. This result validates again the effectiveness of the RSC model.

Overall, these six categories of experiments verify that the SR-based biometric quality assessment and the associated adaptive multimodal recognition are highly effective. The RSC model is a beneficial complementary to the quality-based adaptive multimodal recognition.

5. Conclusions

Multimodal biometric systems are believed to improve the recognition accuracy and robustness by integrating evidence

presented by multiple biometric modalities. In this paper, we gave an overview of multimodal biometrics using the face and ear. All the approaches are classified according to their fusion methodologies. Many multimodal systems have shown the feasibility and advantages of combining the ear and the face. It is a promising way to develop more accurate multimodal systems that are able to recognize a person contactlessly/nonintrusively.

Nevertheless, when the face or the ear suffers from data degeneration, if the fusion rule is fixed or with inferior flexibility, a multimodal system may perform worse than the unimodal system using only the modality with a better quality sample. The biometric quality-based adaptive fusion is an avenue to address this issue. In this paper, we particularly emphasized on a quality-based adaptive multimodal identification system, which adopts a general biometric quality assessment method and dynamically integrates the face and ear via sparse representation. We refined and gave more details about the quality measure and the related fusion weights selection. Moreover, for a more thorough evaluation, we extended the experiments by using more datasets and five types of image degeneration.

Our experimental results demonstrate that the sparse representation-based biometric quality measure is able to mirror the utilities of the face and ear image degenerated by pose, expression, illumination variations, facial disguise such as sunglasses and a scarf, random pixel corruption, and random block occlusion. The quality-based adaptive multimodal method achieves a striking robustness to various types of unimodal corruption/occlusion. Even when the face or ear image suffers from 100% random pixel corruption or random block occlusion, it can still achieve the comparable performance to the unimodal recognition with the ear or the face alone. It can also tolerate a high level of simultaneous face and ear degeneration. In the future, biometric quality assessment and quality-based adaptive multimodal fusion deserve more attention.

Data Availability

No data were used to support this study.

Conflicts of Interest

There are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partly supported by the National Science Foundation of China (61602390, 61860206007, U19A2071, and 615320099), Chinese Ministry of Education (Z2015101), Sichuan Science and Technology Program (2017RZ0009), Education Department of Sichuan Province (15ZB0130), the Fund of Lab of Security Insurance of Cyberspace, Sichuan Province (szjj2015-056), the Xihua University Funds for Young Scholar, and Sichuan Xihua Jiaotong Forensics Center.

References

- [1] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Transactions on Intelligent Systems & Technology*, vol. 7, no. 3, pp. 1–42, 2016.
- [2] Z. Huang, Y. Liu, C. Li, M. Yang, and L. Chen, "A robust face and ear based multimodal biometric system using sparse representation," *Pattern Recognition*, vol. 46, no. 8, pp. 2156–2168, 2013.
- [3] Z. Huang, Y. Liu, X. Li, and J. Li, "An adaptive bimodal recognition framework using sparse coding for face and ear," *Pattern Recognition Letters*, vol. 53, no. 1, pp. 69–76, 2015.
- [4] Z. Huang, Y. Liu, R. Huang, and M. Yang, "Frameworks for multimodal biometric using sparse coding," in *Proceedings of International Conference on Intelligent Science & Intelligent Data Engineering*, pp. 433–440, Nanjing, China, 2012.
- [5] A. F. Abate, M. Nappi, D. Riccio, and G. Tortora, "RBS: A robust bimodal system for face recognition," *International Journal of Software Engineering and Knowledge Engineering*, vol. 17, no. 4, pp. 497–513, 2007.
- [6] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.
- [7] A. K. Jain, P. Flynn, and A. Ross, *Handbook of Biometrics*, Springer, Berlin, Germany, 2007.
- [8] D. Zhang, F. Song, Y. Xu et al., "Advanced Pattern Recognition Technologies with Applications to Biometrics," *IGI Global*, Hershey PA, USA, 2008.
- [9] S. Li and A. K. Jain, *Encyclopedia of Biometrics*, Springer, Berlin, Germany, 2009.
- [10] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, no. 5, pp. 449–480, 2004.
- [11] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1160–1165, 2003.
- [12] A. K. Jain and A. Ross, "Fingerprint mosaicking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4064–4067, Orlando, FL, USA, 2002.
- [13] C. Liu and H. Wechsler, "Face recognition using shape and texture," in *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pp. 23–25, Fort Collins, CO, USA, 1999.
- [14] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu, "Feature fusion: Parallel strategy vs. serial strategy," *Pattern Recognition*, vol. 36, no. 6, pp. 1369–1381, 2003.
- [15] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.
- [16] L. Yuan, Z.-C. Mu, and X.-N. Xu, "Multimodal recognition based on face and ear," *International Conference on IEEE Wavelet Analysis and Pattern Recognition*, vol. 3, pp. 1203–1207, 2007.
- [17] D. Yaman, F. I. Eyiokur, and H. K. Ekenel, "Multimodal age and gender classification using ear and profile face images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, June 2019.

- [18] X. Xu and Z. Mu, "Multimodal recognition using ear and face profile based on CCA," *Application Research of Computers*, vol. 24, no. 11, pp. 312–314, 2007, in Chinese.
- [19] X. Xu and Z. Mu, "Feature fusion method based on KCCA for ear and profile face based multimodal recognition," in *Proceedings of the 2007 IEEE International Conference on Automation and Logistics*, pp. 620–623, Jinan, China, August 2007.
- [20] T. Theoharis, G. Passalis, G. Toderici, and I. A. Kakadiaris, "Unified 3D face and ear recognition using wavelets on geometry images," *Pattern Recognition*, vol. 41, no. 3, pp. 796–804, 2008.
- [21] G. S. Badrinath and P. Gupta, "Feature level fused ear biometric system," in *Proceedings of the Seventh International Conference on Advances in Pattern Recognition (ICAPR 2009)*, pp. 4–6, Kolkata, India, February 2009.
- [22] X. Xu and Z. Mu, "Multimodal recognition based on fusion of ear and profile face," in *Proceedings of the Fourth International Conference on Image and Graphics*, pp. 598–603, Chengdu, China, 2007.
- [23] P. Yan, "Ear biometrics in human identification," Ph.D. thesis, University of Notre Dame, Notre Dame, Indiana, 2006.
- [24] M. H. Mahoor, S. Cadavid, and M. Abdel-Mottaleb, "Multimodal ear and face modeling and recognition," in *Proceedings of the 16th IEEE International Conference on Image Processing*, pp. 4137–4140, Cairo, Egypt, November 2009.
- [25] S. Cadavid, M. H. Mahoor, and M. Abdel-Mottaleb, "Multimodal biometric modeling and recognition of the human face and ear," in *Proceedings of the IEEE International Workshop on Safety, Security & Rescue Robotics (SSRR)*, pp. 1–6, Denver, CO, USA, 2009.
- [26] S. M. S. Islam, M. Bennamoun, A. S. Mian et al., "Score level fusion of ear and face local 3D features for fast and expression-invariant human recognition," *Image Analysis and Recognition*, pp. 387–396, Springer, Berlin, Germany, 2009.
- [27] A. A. Darwish, R. A. Elghafar, and A. F. Ali, "Multimodal face and ear images," *Journal of Computer Science*, vol. 5, no. 5, pp. 374–379, 2009.
- [28] D. R. Kisku, P. Gupta, J. K. Sing et al., "Fusion of gaussian mixture densities for face and ear biometrics using support vector machines," *Future Generation Information Technology*, pp. 344–351, Springer, Berlin, Germany, 2010.
- [29] R. Raghavendra, R. Ashok, and G. Hemantha Kumar, "Multimodal biometric score fusion using gaussian mixture model and Monte Carlo method," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 771–782, 2010.
- [30] A. P. Yazdanpanah, K. Faez, and R. Amirfattahi, "Multimodal biometric system using face, ear and gait biometrics," in *Proceedings of the 10th International Conference on Information Sciences Signal Processing & their Applications*, pp. 251–254, Kuala Lumpur, Malaysia, 2010.
- [31] Z. Huang, Y. Liu, and X. Wang, "Study on sparse representation based classification for biometric verification," 2015.
- [32] Z. Huang, Z. Feng, J. Kittler, and Y. Liu, "Improve the spoofing resistance of multimodal verification with representation-based measures," in *Proceedings of the First Chinese Conference on Pattern Recognition and Computer Vision*, Guangzhou, China, 2018.
- [33] M. M. Monwar and M. Gavrilova, "FES: A system for combining face, ear and signature biometrics using rank level fusion," in *Proceedings of the Fifth International Conference on Information Technology: New Generations*, pp. 922–927, Las Vegas, NV, USA, April 2008.
- [34] M. M. Monwar and M. L. Gavrilova, "Multimodal biometric system using rank-level fusion approach," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 4, pp. 867–878, 2009.
- [35] M. M. Monwar and M. Gavrilova, "Integrating monomodal biometric matchers through logistic regression rank aggregation approach," in *Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop*, pp. 1–7, Washington, DC, USA, October 2008.
- [36] A. Kumar, M. Hanmandlu, and S. Vasikarla, "Rank level integration of face based biometrics," in *Proceedings of the 11th International Conference on Information Technology: New Generations*, pp. 36–41, Las Vegas, NV, USA, 2012.
- [37] M. M. Rahman and S. Ishikawa, "Proposing a passive biometric system for robotic vision," in *Proceedings of the Tenth International Symposium on Artificial Life and Robotics*, pp. 4–6, Oita, Japan, 2005.
- [38] D. R. Kisku, P. Gupta, H. Mehrotra, and J. K. Sing, "Multimodal belief fusion for face and ear biometrics," *Intelligent Information Management*, vol. 1, no. 3, pp. 166–171, 2009.
- [39] N. B. Boodoo and R. K. Subramanian, "Robust multi-biometric recognition using face and ear images," 2009, <http://arxiv.org/abs/0912.0955>.
- [40] E. Candes, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians*, Madrid, Spain, August 2006.
- [41] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
- [42] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pp. 625–632, Colorado Springs, CO, USA, June 2011.
- [43] A. Hicklin and R. Khanna, *The Role of Data Quality in Biometric Systems*, White Birch Paper Co, Greenwich, CT, USA, 2006.
- [44] F. A. Fernández, "Biometric sample quality and its application to multimodal authentication systems," Ph.D. thesis, Universidad Politécnica de Madrid, Madrid, Spain, 2008.
- [45] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.
- [46] A. M. Martinez and R. Benavente, "The AR face database," CVC Technical Report, Purdue University, West Lafayette, Indiana, 1998.
- [47] USTB Ear Database, University of Science & Technology Beijing, Beijing, China, 2011.
- [48] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1 -regularized least squares," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.
- [49] A. S. Georgiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.
- [50] Georgia Tech Face Database, http://www.anefian.com/research/face_reco.htm, 2011.