Learning pairwise SVM on hierarchical deep features for ear recognition

Ibrahim Omara1,2, Xiaohu Wu1, Hongzhong Zhang1, Yong Du3, Wangmeng Zuo1,2,3
1 School of Computer Science and Technology, Harbin Institute of Technology (HIT), Harbin, People’s Republic of China
2 Department of Mathematics, Faculty of Science, Menoufia University, Shebin El-koom, Egypt
3 Department of Electrical and Information Engineering, Northeast Agricultural University, Harbin, People’s Republic of China
E-mail: cswmzuo@gmail.com

Abstract: Convolutional neural networks (CNNs)-based deep features have been demonstrated with remarkable performance in various vision tasks, such as image classification and face verification. Compared with the hand-crafted descriptors, deep features exhibit more powerful representation ability. Typically, higher layer features contain more semantic information, while lower layer features can provide more low-level description. In addition, it turns out that the fusion of different layer features will lead to superior performance. Here, we propose a novel approach for human ear identification by combining hierarchical deep features. First, hierarchical deep features are extracted from ear images using CNN pre-trained on large-scale data set. To enhance the feature representation and reduce the high dimension of deep features, the discriminant correlation analysis (DCA) is adopted for fusing deep features from different layers for further improvement. Owing to the lack of ear images per person, the authors propose to transform the ear identification problem to the binary classification by composing pairwise samples and resolve it with the pairwise support vector machine (SVM). Experiments are conducted on four public databases: USTB I, USTB II, IIT Delhi I, and IIT Delhi II. The proposed method achieves promising recognition rate and exhibits decent performance compared with the state-of-the-art methods.

1 Introduction

Biometric systems have attracted much attention on various applications, especially in the fields of security and forensics [1]. In general, biometric traits can be grouped into two categories, physiological and behavioural ones. Physiological traits include face, fingerprint, ear, hand geometry, finger vein, and iris etc., while behavioural traits involve gait, keystroke dynamics, signature etc. In the last few decades, varieties of biometric recognition methods have been studied on face, fingerprint, palmprint, and iris [2, 3, 4]. Recently, ear print has received considerable research interests in biometric community due to its several prominent advantages [5]. Human ears are large and visible for acquisition, stable through age and expressions, and are different for identical twins. It satisfies all the required biometric characteristics such as universality, uniqueness, permanence, and measurability. In [6], Iannarelli has shown that there are no two ears having the same helix. In Fig. 1, one example of ear image is provided for better understanding the structure of human ear, which can be divided into the outer ear shape and inner ear shape, and includes the inner and outer helix, antihelix, tragus, antitragus, concha, crus helix, and lobe.

Despite their intensive applications in computer vision, hand-crafted features (e.g. the scale-invariant feature transform [7]) and the visual descriptors such as bag of visual words [8] usually are limited in discrimination and generalisation. Recently, driven by the advances in convolutional neural networks (CNNs), deep features have been drawing much attention due to its excellent representation and discrimination ability, which have been widely adopted and demonstrated impressive performance. Typically, the top layers of CNN contain more semantic information and describe the global feature of the images, while the intermediate layers describe the local features and the bottom layers contain more low-level information for the description of textures and edges. Compared with the handcrafted descriptors, deep features can provide more discriminative information and lead to better performance, especially in object detection, object tracking, and image classification [9, 10]. Moreover, CNNs have been studied intensively in various fields, such as face and handwritten digit recognition, object detection, medical image analysis [11, 12], and eye detection [13]. However, less effort has been paid on the deep feature representation of ear images for ear recognition.

On the other hand, fusion algorithms have been widely adopted in pattern recognition. Different levels of fusion methods have been studied, such as pixel level [14], feature level [14], and decision level fusion [15]. Among these methods, feature level fusion has been intensively studied [16]. It is noted that different features may provide descriptions of the object from different aspects and contain complementary information. By combining these features, we can not only keep the effective discriminative information of multi-features, but also eliminate the redundant information to certain degree. Obviously, all these benefit to the classification performance. The typical feature fusion methods include canonical correlation analysis (CCA)-based fusion [17], parallel feature fusion [18], and serial feature fusion [16]. Liu et al. [16] concatenate different feature vectors into one, and Yang et al. [18] combine two feature vectors into a complex vector. The CCA-based fusion method aims to maximise the correlation between two sets of features and has been widely used in feature fusion [19]. Instead of CCA, in this work, we use the discriminant correlation analysis (DCA) [20] for fusing different layer of features of ear image due to its simplicity and efficiency.

Fig. 1 Anatomy of the ear
Matching is processed after feature extraction, which aims to generate score for indicating the similarity of two ear images. In the nearest neighbour (NN) classification, similarity is measured based on some distance metric, such as Euclidean distance [21, 22] and Hamming distance [23]. As for ear identification, discriminative classifiers, such as the support vector machine (SVM), can be adopted for classification [24]. However, due to the number of ear images per person is very limited (e.g. 3–5 for most public data sets), SVM cannot achieve the desired performance. Thus, we propose to apply pairwise SVM [25] to solve the problem. Pairwise SVM provides an extension of the binary SVM classifier to multi-class classification, which has been demonstrated to get excellent results on the classification problem.

To sum up, we present an ear identification method based on hierarchical CNN features, which is an extension of our previous work [26]. Specifically, we focus on the feature extraction and classification modules. Compared with [26], DCA is adopted to combine hierarchical feature from different layers, and more experiments are conducted on four public ear databases. The contributions can be summarised as follows:

- Benefiting from the powerful representation ability of deep features, we propose to represent the ear images with deep features for ear recognition.
- DCA [20] is utilised to perform dimensionality reduction and fuse different layers of features for better representation of the ear images. Furthermore, to handle the ear image deficiency in training, we suggest using the pairwise SVM to transform ear identification into a binary classification task.
- Extensive experiments are conducted to evaluate the proposed method on four public databases, i.e. USTB databases collections I and II, and IIT Delhi databases collections I and II. The results show that the proposed method performs favourably over the state-of-the-art methods.

The rest of the paper is organised as follows: Section 2 introduces the related work about the feature extraction and matching methods for ear recognition. Section 3 presents the proposed method based on deep features, discriminative correlation analysis, and pairwise SVM. Section 4 reports the experimental results. Finally, Section 5 draws the conclusion of the paper.

2 Related work

In this section, we separately discuss the methods for ear feature extraction, feature fusion, matching, and classification.

2.1 Ear feature extraction

In this work, we use deep CNN for ear feature extraction. Instead of deep features, many hand-crafted features, including both holistic [27] and local [7, 28–30] ones, have been suggested for ear images. Compared with hand-crafted features, deep CNN features are learned in a data-driven manner, and have proven its effectiveness in many vision tasks. Moreover, it has been demonstrated that the features learned from the CNNs, such as AlexNet [31] and VGG-Net [32] pre-trained on ImageNet, can be well transferred to biometric recognition tasks, such as face recognition [33], iris recognition [34], signature [35], and finger-vein identification [36]. In this work, we show that the hierarchical CNN features can also be extended to ear feature representation with promising performance.

2.2 Feature fusion

Different layers of CNN features may contain complementary information for ear recognition. Feature fusion can thus be explored to improve the representation ability of deep CNN features from different layers. For general feature fusion, Yang et al. [18] propose a parallel scheme which makes two sets of data to form a complex feature vector, and then perform the complex linear projection to improve classification accuracy. Sun et al. [17] adopt the idea of CCA and propose a feature fusion method to maximise the correlation between two sets of features. Based on CCA, Haghight et al. [20] propose the DCA for multi-modal biometrics feature fusion, which considers the class associations in feature sets, eliminates the between-class correlations, and restricts the correlations to be within classes. In this paper, we exploit the DCA fusion algorithm to fuse different layers of features extracted from VGG-Net [32] to further improve the discriminability and robustness of the representation for ear recognition.

2.3 Ear matching and classification

After feature extraction, a matching method or classifier is needed to recognise the test ear image. Commonly adopted method for matching criterion is the distance metric, such as Euclidean distance and Hamming distance [21–23]. SVM has also been widely adopted for ear classification as a classifier based on principal component analysis (PCA) and independent component analysis (ICA) features in [37]. However, due to the lack of training images and the multi-class characteristics for ear recognition, SVM may not lead to satisfying performance. Pairwise SVM [25] extends multi-class SVM to binary SVM from the similarity matching perspective and exhibits better interclass generalisation performance, which can be applied to better handle the multiple classes of ear image database. Therefore, in this paper, we apply pairwise SVM for ear recognition.

3 Proposed method

In this section, we describe the proposed method for ear identification in details. Fig. 2 illustrates the framework of the proposed method, which includes the deep features extraction from VGG-M [38], the DCA [20] algorithm for fusing different layers deep features, and the pairwise SVM classifier for ear recognition. In the following subsections, we will describe each part of the proposed method.

3.1 Deep features extraction

In this work, VGG-M [38] pre-trained on the ILSVRC data set is employed for feature extraction based on the following considerations. First, the ear image usually has limited size, which implies that a proper depth of deep model should be considered. For ear identification, the depth of VGG-M is sufficient to extract the local and global features of ear image. Second, considering the computational budget and memory storage, those very deep models, such as VGG-16, VGG-19 [32], and ResNet [39], usually require more resources and may be difficult to be deployed in practical biometric recognition systems. Finally, due to its excellent performance and promising transferability, VGG-M has been widely adopted in various vision tasks, such as multi-biometric recognition [40]. Moreover, we conduct experiments to compare the performance of the VGG-M and VGG-16. It can be found that VGG-M is on par with VGG-16 and the fine-tuned VGG-M in terms of recognition accuracy.

As illustrated in Fig. 2, VGG-M contains eight layers, where the first five are convolution layers, and the remaining three are fully connected layers. Specifically, the first and second convolution layers are followed by response normalisation layer, and max-pooling layers are stacked upon the response normalisation layers as well as the fifth convolution layer. Furthermore, the rectified linearity unit (ReLU) is applied to the output of each convolution and fully connected layer. Table 1 summarises the VGG-M architecture. Owing to the fixed-size input (e.g. 224 × 224 × 3) of VGG-M, we copy the grey image to three channels, and resize each ear image to 224 × 224 in experiments.

For better understanding the deep features of ear images, Fig. 3 visualises several outputs of different convolution layers corresponding to four ear images from the USTB I, II, and IIT Delhi I, II data sets, respectively. We can see that these features do capture some meaningful information about ears. Specifically, the bottom features contain more contour and edges information. Owing to the property of CNN, we know that the top layer of features convey more semantic information to describe the global feature of ear images, while the middle and bottom layers of features contain more local features, such as edges and texture.
details. In Section 4, we conduct evaluation experiments on the convolution layers and the fully connected layers separately to analyse the effect of features from different layers for ear recognition. Note that different layers of deep features convey complimentary information of ear image. We will perform the feature fusion with the DCA algorithm for enhanced ear representation.

3.2 Feature fusion with DCA

DCA [20] is proposed for feature level fusion with low computational complexity. It aims to maximise the pairwise correlation across the two different feature sets, while eliminating the between-class correlations and restricting the correlations to be within-class. DCA can be used to fuse different feature vectors that extracted from multiple traits or from a single trait. Since different layers of features extracted from VGG-M have its own characteristics, we adopt DCA algorithm to fuse different features of ear images to further enhance the representation ability.

Suppose that there are \( n \) ear images which belong to \( c \) subjects. Denote by \( \mathbf{s} = \{(x_i, y_i), (x_i, y_i), \ldots, (x_i, y_i)\} \) the training set, where \( x_i \) stands for the \( i \)th sample and \( y_i \in \{1, 2, \ldots, c\} \) stands for the corresponding class label. \( X \) and \( X_i \) denote the training set matrix composed by different layers of features extracted from VGG-M, respectively. However, the ear images have similar appearance and the number of training images is limited to 3–5 for each subject, further increasing the difficulty of training a discriminative classifier. For the purpose of distinguishing the ear images from different subject, DCA integrates the class associations into the correlation analysis, which motivates the transform of features based on the between-class scatter matrix

\[
\mathbf{S} = \sum_{i=1}^{c} n_i (\bar{x}_i - \bar{x}) (\bar{x}_i - \bar{x})^\top,
\]

where \( n_i \) denotes the number of ear images belonging to the \( i \)th subject, \( \bar{x}_i = (1/n_i) \sum x_i \) denotes the centre ear image of the \( i \)th subject, and \( \bar{x} = (1/n) \sum x \) denotes the centre of all ear samples. However, deep features always have the property of high dimension, which leads to the high computational complexity. According to [41], instead of decomposing the between-class scatter matrix \( \mathbf{S} \), the covariance matrix \( \mathbf{C} = \Phi \Phi^\top \) (\( \Phi = (\sqrt{n_1} (\bar{x}_1 - \bar{x}), \ldots, \sqrt{n_c} (\bar{x}_c - \bar{x})) \)) can be decomposed due to its semi-positive definite (SPD) property:

\[
\mathbf{C} = \mathbf{U} \Lambda \mathbf{U}^\top, \Lambda = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_r), \lambda_1 \geq \lambda_2, \ldots \geq \lambda_r \tag{1}
\]

where \( \lambda_i \) is the \( i \)th eigenvalue of \( \mathbf{C} \), and the \( i \)th column of \( \mathbf{U} \) is the corresponding eigenvector. Considering the first largest \( r \) non-zero eigenvalues, and their corresponding eigenvectors \( \mathbf{U}_r \), the
transformation matrix is defined as $P = \Phi U A^{-1/2}$. Typically, for feature set $X_i$, it is projected to $Z_i = P_i X_i$ and the feature set $X_i$ is transformed to $Z_i = P_i X_i$ correspondingly.

Deep feature vector generally is of high dimension, which implies that there may contain redundant features. One conventional strategy is to use PCA whitening [42] for dimensionality reduction. Note that the goal of DCA is to fuse different feature sets. Thus, after PCA whitening, DCA further suggests to make the features in one set have non-zero correlation only with their corresponding features in the other set. In the implementation, we adopt an adaptive method to set the dimension of PCA by keeping the ratio of the sum of the selected eigenvalues and the sum of all eigenvalues $>0.95$. Based on the transformed feature sets $Z_i$ and $Z_j$, the between-set covariance matrix is defined as $S_b = Z_i Z_j^T$. Owing to the SPD property, singular value decomposition (SVD) on $S_b$ is $S_b = V \Sigma V^T$. Then the transform matrix is defined as $T = V \Sigma^{-1/2} Z_i$ and $Z_i$ are further transformed to $X_i = T^T Z_i$ and $X_i = T^T Z_j$. Finally, the fused feature by DCA is defined as the summation of the transformed feature vectors: $X = X_1 + X_2$ for the sake of low dimension. The detailed process can be found in Algorithm 1.

Algorithm 1: The DCA algorithm for deep features fusion

**Input:** $X_i, X_j$

**Output:** $X$

1: Compute the mean vector for each class: $\bar{x}_i = (1/n_i) \sum_{i=1}^{n_i} x_{ij}$, and the mean of training data: $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$, based on.

2: Compute the covariance matrix: $C = \Phi^T \Phi$, where $\Phi = (\sqrt{n_1}(\bar{x}_1 - \bar{x}), \ldots, \sqrt{n_2}(\bar{x}_2 - \bar{x}))$.

3: Compute the SVD of $C$: $C = U \Sigma U^T$, $U = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) (\lambda_1 \geq \lambda_2 \ldots \geq \lambda_d)$.

4: Define the transform matrix: $P = \Phi U A^{-1/2}$, (Lines 1–4 provide the same process performed on $X_i$ and $X_j$ to get their corresponding transform matrix separately.)

5: Compute the transformed data: $Z_i = P_i X_i$ and $Z_j = P_j X_j$.

6: Compute the between-set covariance matrix of the transformed feature set: $S_b = Z_i Z_j^T$.

7: Compute the SVD of $S_b$: $S_b = V \Sigma V^T$.

8: Define the transform matrix: $T = V \Sigma^{-1/2}$.

9: Compute the transformed data: $X_i = T^T Z_i$ and $X_j = T^T Z_j$.

10: The final transformed data: $X = X_1 + X_2$.

3.3 Pairwise SVM

Given an ear image, the identification task aims to determine which person it belongs to. In this work, we transform the identification task to the binary classification problem, and introduce the pairwise SVM [25] to resolve it. Pairwise SVM [25] relies on a pair of input samples to predict whether they are from the same subject or not. Denote by the ear images set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where $x_1$ presents the feature representation of one ear image and $y_1$ stands for which person it belongs to. First, pairwise inputs are required consisting of the positive pair and negative pair samples. For each subject, we compose pairwise example as $[x_i, x_j]$ from the training images set. If $x_i$ and $x_j$ belong to the same subject, then $[x_i, x_j]$ is defined as the positive pair with label $p_{ij} = +1$, which is further extended to positive pairs $[x_i, x_j], [x_j, x_i]$, and $[x_i, x_j]$. On the contrary, if $x_i$ and $x_j$ belong to the different persons, then $[x_i, x_j]$ is assigned with negative label $p_{ij} = -1$, with pairwise sample $[x_i, x_j]$. Owing to the lack of ear images of each subject, there may exist the problem of unbalanced positive and negative pairs. Therefore, we take several strategies to relieve this issue. First, we extend $(x_i, x_j)$ to $(x_i, x_j)$ for producing more positive pairs, which also satisfies the symmetric testing set constraint required by pairwise SVM. Second, we replicate the original positive pairs to generate more positive data. In addition, we integrate KNN algorithm into the composition of negative pairs. Instead of using all negative pairs or selecting some of them randomly, we choose the $k$ nearest ear images from the different persons to form the set of negative pairs. Furthermore, to alleviate the class imbalance, we control the ratio of positive and negative pairs manually during the training of pairwise SVM. Through the strategies above, we finally balance the positive and negative pairs and also experimentally evaluate the effect of the ratio of positive and negative pairs.

Benefited from the kernel trick $k(x, z) = \langle \phi(x), \phi(z) \rangle$, $\phi: X \rightarrow \mathcal{R}$, the pairwise kernels are defined as:

$$K: (X \times X) \times (X \times X) \rightarrow \mathbb{R}$$

and the decision function should be:

$$f(x_i, x_j) = \sum_{q \in I, \eta \neq I} \alpha_{q \eta} K(x_q, x_{\eta}) + b$$

where $I \subseteq n \times n$. In pairwise SVM, the goal is to decide whether the examples of a pair $(x_i, x_j)$ belong to the same class or not, which implies the symmetric property of the decision function: $f(x_i, x_j) = f(x_j, x_i)$. Thus, the balanced kernel is motivated to keep the symmetric of the decision function. More details and discussions about the balanced kernels can be found in [25].
addition to the balanced kernel, a symmetric pairwise decision function also can be achieved by symmetric training sets. Instead of exploiting the different balanced kernels, e.g. the metric learning pairwise kernel $K_{ML}$, the tensor learning pairwise kernel $K_{TM}$, we adopt the symmetric training sets for learning the pairwise SVM. One reason is to address the lack of training samples and the requirement on symmetric training set by exploiting the positive data augmentation strategy discussed above. Also the symmetric training set makes it easy to utilise the off-the-shelf SVM toolbox. The LibSVM toolbox is utilised for implementing pairwise SVM.

To describe the identification procedure, we take the USTB I as an example with two ear images of each subject for training. Suppose that the training set is denoted as $\mathcal{Z} = \{x_1^i, x_2^i, \ldots, x_c^i\}$, where $c$ means the number of subjects and $x_i^c$ presents the $i$th ear image that belongs to the $i$th subject. The test set can be denoted as $\mathcal{Z} = \{z_1, \ldots, z_c\}$. Given the pairwise SVM model, we compose pairwise samples \{(z_1, x_1^i), (z_2, x_2^i), \ldots, (z_c, x_c^i)\} for each test data $z_i$ where $i \in \{1, 2, \ldots, c\}$. According to the decision values of the pairwise SVM, we choose the label of the one with the highest positive score as the predicted label for the test sample $z_i$.

4 Experimental results

4.1 Data sets and evaluation metrics

The experiments are performed on four publicly available ear databases, i.e. USTB collections I, II, and IIT Delhi collections I, II. The USTB databases are collected from the students and teachers at the Department of Information Engineering, University of Science and Technology of Beijing (USTB) [43]. The IIT Delhi databases are acquired from the students and staffs at IIT Delhi, New Delhi [44], and all ear image databases have been collected from the volunteers during October 2006-June 2007 in the indoor environment. Fig. 4 shows the original ear images from these four databases. A brief summary of the databases is given as follows.

**USTB I Database:** It consists of 180 images of 60 subjects, among which each subject has 3 images. The images are taken under standard illumination and trivial rotation, and all images have the same size 150 × 80 pixels.

**USTB II Database:** It comprises 308 images of 77 subjects, which are taken under illumination and angle change condition. Each subject has four images: the first image is the frontal ear image under standard illumination, the second and the third images are obtained with the rotation of +30° and −30°, respectively. Also, the fourth image is taken under weak illumination. All images in USTB II database have been acquired during November 2003 to January 2004 and have the same size of 400 × 300.

**IIT Delhi I Database:** This database contains 493 images of 125 subjects. Every subject includes at least three images with different pose and lighting condition, and the image has the size of 272 × 204.

**IIT Delhi II Database:** It includes images of 221 persons, among which each subject has at least three images. The images are automatically segmented and normalised for 125 subjects from IIT Delhi I database and other set of subjects, and all the images have the same size of 180 × 50.

**Experimental settings:** In our experiments, we set the default $k$ to be 5 to address the class imbalance problem. The trade-off C in pairwise SVM is determined by 10-fold inner cross-validation from \{2^{\text{min}} \cdot \text{step}^{\text{max}}\}, with $c_{\text{min}} = -5$, $c_{\text{step}} = 1$, and $c_{\text{max}} = 15$. In the ablative experiments, we randomly select $r = 2$ ear images per subject to constitute the training set, and use the remaining images as the test set. To avoid the effect of training/testing partitions, the average performance metrics are adopted by running such experiment 20 times. When compared with the competing methods, we adopt the widely used protocol in the literature for fair comparison.

**Evaluation metrics:** In the experiments, we consider both the close set and the open set identification setting [45], and adopt the following evaluation metrics:

**Closed-set identification:** Given a test image $x_{i}^{\text{test}}$, pairwise SVM is used to compute its similarity score to each training image. By sorting the similarity scores, we can obtain the rank of the ground-truth subject, i.e. rank($x_{i}^{\text{train}}$). The rank-$k$ recognition rate [45, 46] is then defined as

$$\text{Acc}(k) = \frac{|\{x_{i}^{\text{test}} | \text{rank}(x_{i}^{\text{train}}) \leq k\}|}{N_{\text{test}}},$$

where $|\cdot|$ denotes the cardinality of a set, and $N_{\text{test}}$ the number of test images. In our experiments, the closed set identification performance is measured by the rank-1 recognition rate Acc(1).

**Open-set identification:** Given a test image $x_{i}^{\text{test}}$, the system first performs a detection subtask to determine whether it belongs to the set of subjects by a decision threshold $r$. After that, it further conducts an identification subtask to decide the subject of $x_{i}^{\text{test}}$. Following [45, 47], we define the detection and identification rate (DIR) at $r$ as

$$\text{DIR}(r) = \frac{|\{x_{i}^{\text{test}} | \text{rank}(x_{i}^{\text{train}}) = 1, f(x_{i}^{\text{test}}, x_{j}^{\text{train}}) \geq r, \forall j\}|}{N_{\text{test}}},$$

where $x_{i}^{\text{train}}$ denotes the $j$th training image, and the similarity function $f(x_{i}^{\text{test}}, x_{j}^{\text{train}})$ is defined in (3). Owing to that the non-registered test images are unavailable, in our experiments, we slightly modify the definition of false alarm rate (FAR) in [45, 47] as

$$\text{FAR}(r) = \frac{|\{x_{i}^{\text{test}} | \max_{x_{j}^{\text{train}}} f(x_{i}^{\text{test}}, x_{j}^{\text{train}}) \geq r\}|}{N_{\text{test}}},$$

where ID(·) denotes the subject of the sample. The ROC curve for identification can then be obtained as the plot of DIR versus FAR by varying the decision threshold $r$, namely DIR–FAR curve. Analogous to [46], we also report the rank-1 recognition rate at 1% FAR (i.e. DIR at 1%).

To evaluate the proposed approach, we further conduct comparison experiments with several state-of-the-art methods. For comprehensive evaluation, we consider three scenarios in experiments, and compare the proposed method with the competing ones reported with the similar settings. In scenario 1, one image of each subject is taken for training, and the remaining images are for testing. In scenarios 2 and 3, about two and three images per subject are selected for training, respectively, and the remaining images are used for testing.
Specifically, the learning rate starts from 0.01 and is divided by 10 in Fig. 4, the image in the USTB I contains the ear only, which after every 40 epochs. The training is terminated at the 120 epochs, connected features on all CNN models without positive pairs that the convolutional features show its superiority to the fully deep features, we evaluate three CNN models, including the VGG-M, VGG-16, and the fine-tuned VGG-M with the training images. In the fine-tuning stage of VGG-M, we use the softmax loss. More negative samples will decrease the recognition rate. However, from Fig. 7, we note that a little more negative data will improve the performance on the USTB I data set. With positive pairs replication, the recognition rate on IIT Delhi I has achieved an extreme experiment on USTB I with the ratio of 1:33 and 1:36. The recognition rates are 56.67 and 36.67%, respectively, demonstrating the necessity of controlling the ratio. We also note that the replication of positive pairs has larger influence on the recognition rate due to the data deficiency. A balance between the positive and negative pairs benefits to the performance, and the positive data augmentation is also helpful according to the experimental results. In the following experiments, we fix the ratio to 1:3 and augment the positive pairs with two replications if not explicitly emphasised.

4.2 Deep features analysis

To analyse the effect of the different layers of features, we conduct experiments on the USTB I and IIT Delhi I data sets by using two images per subject for training. To evaluate the generalisation of deep features, we evaluate three CNN models, including the VGG-M, VGG-16, and the fine-tuned VGG-M with the training images. In the fine-tuning stage of VGG-M, we use the softmax loss. Specifically, the learning rate starts from 0.01 and is divided by 10 after every 40 epochs. The training is terminated at the 120 epochs, and the batch size is set as 16. Furthermore, data augmentation, including random flipping, shifting, and rotation, is adopted during training. Considering the computational efficiency and memory storage we employ the convolutional features \{conv3, conv4, conv5\} and fully connected features \{fc6, fc7, fc8\} in VGG-M and the fine-tuned VGG-M. As for VGG-16, the deep features from \{conv3, conv4, conv5, conv6, fc6, fc7, fc8\} are utilised. For simplicity, we unify them to \{conv3, conv4, conv5\} and \{fc6, fc7, fc8\}.

Fig. 5 shows the ear recognition performance by using the different layers of features and the CNN models. It can be observed that the convolutional features show its superiority to the fully connected features on all CNN models without positive pairs augmentation. We extend the experiments to be with positive pairs augmentation. By replicating the original positive pairs twice, we increase the number of the positive pairs three times. Furthermore, the ratio of positive and negative data is fixed to 1:3. It is obvious that the positive data augmentation leads to excellent performance on each layer. However, there are still test images not correctly predicted using the fully connected features. In particular, a recognition rate of 80.00% around is obtained by using the fc8 layer features on the USTB I data set, which means that there are about ten test images misclassified by pairwise SVM. The fully connected features of VGG-16 perform better than the VGG-M on USTB I, while worse on the IIT Delhi I data set, which should owe to the characteristics of ear images in different data sets. As shown in Fig. 4, the image in the USTB I contains the ear only, which means that all values in the fully connected layers correspond to the ear, while for the IIT Delhi I, some background information is included and may distract the classifier. With the convolutional features, VGG-16 performs favourably against VGG-M on both data sets. Furthermore, it can be seen that the pre-trained VGG-M performs comparably to the fine-tuned VGG-M model. The results validate the generalisation of the deep models pre-trained on large-scale data set, which can be well transferred to extract deep features for ear images. Considering the comparative ear recognition rate and the feature extraction efficiency discussed in Section 4.7, VGG-M can serve as a suitable choice for ear image feature extraction.

4.3 Data imbalance analysis

From Fig. 5, one can note that the fully connected features decrease the recognition rate on the USTB I data set. On IIT Delhi I, the features from fc8 also lead to worse performance, while better recognition rate can be obtained by using the convolutional features. However, the higher dimension of the convolutional features may affect the efficiency. Thus, we hope to further improve the performance with the fully connected features. Under the assumption that compared with the convolutional features, the data imbalance between the positive and negative pairs has larger influence on that by using the fully connected layer features. To address the class imbalance problem, we propose to replicate the positive pairs and integrate the KNN into the selection of negative pairs as mentioned in Section 3.3. We further conduct experiments to evaluate the effects of the positive pairs replication and the ratio of positive and negative pairs. Without positive pairs replication, we perform experiments with different ratios of positive and negative pairs on the USTB I and IIT Delhi I data sets. From Fig. 6, we note the proper ratio should be in the range of 1:1–1:7. More negative samples will decrease the recognition rate. However, from Fig. 7, we note that a little more negative data will improve the performance on the USTB I data set. With positive pairs replication, the recognition rate on IIT Delhi I has achieved excellent performance and is more robust to the ratio. We perform an extreme experiment on USTB I with the ratio of 1:33 and 1:36. The recognition rates are 56.67 and 36.67%, respectively, demonstrating the necessity of controlling the ratio. We also note that the replication of positive pairs has larger influence on the recognition rate. Therefore, we design experiments to evaluate the positive pair replication with the fixed ratio of 1:3. The results in Fig. 8 validate the promising performance of the positive pairs augmentation both on the USTB I and IIT Delhi I data sets.

To sum up, the class imbalance has serious effect on the recognition rate due to the data deficiency. A balance between the positive and negative pairs benefits to the performance, and the positive data augmentation is also helpful according to the experimental results. In the following experiments, we fix the ratio to 1:3 and augment the positive pairs with two replications if not explicitly emphasised.
4.4 Feature fusion analysis

From Figs. 5 and 8, we have the following two observations. First, convolutional features outperform fully connected features, and a single layer of convolutional features without positive pair replication can achieve better performance than fully connected features. Second, the positive pair replication improves the recognition rate largely for fully connected features. However, the improvement by positive pair replication depends on the data set, and is complexly related to the ratio of positive and negative pairs which should be evaluated manually. For further enhancing ear image representation, we employ the DCA algorithm for feature fusion to improve the recognition rate. To analyse the effect of feature fusion, experiments are conducted on USTB I and IIT Delhi I data sets. For fair comparison, the evaluation experiments on the feature fusion by DCA algorithm are performed without positive pair replication.

It is well known that different layers of features contain complementary information. In particular, the lower layers of convolutional feature contain more low-level information such as textures and edges, while the higher layers of fully connected features convey more semantic information. With two images of each subject for training, we fuse any two layers from the convolutional features \{conv3, conv4, conv5\} and the fully connected layers of features \{fc6, fc7, fc8\}. According to the previous evaluation, we fix the ratio of the positive and negative pairs to 1:3. The results are listed in Table 2. One can see that any two layers of feature fusion improve the recognition rate. While single layer of convolutional features achieves considerable performance, the fully connected layers features benefit more from the fusion scheme by DCA. Typically, the feature fusion gains an obvious improvement in comparison with the use of single fully connected layer of features on both the USTB I and IIT Delhi I data sets. In the following, we will utilise the fusion of features from the fc6 and fc7 layers.

4.5 Occlusion analysis

In real-world applications, the ear images are likely to be occluded by hair or other distractors. To evaluate the robustness to occlusion, we conduct experiments with random occlusion on USTB I and IIT Delhi I data sets. According to [48], we add random occlusion with
different percentages, such as 10, 20, 30, and 40% on the ear images. We perform three groups of experiments based on the feature fusion of fc6 and fc7 by DCA, occlusion on the training images, occlusion on the test images, and occlusion on both of them. As shown in Fig. 9, the proposed method is quite robust to random occlusion. With 10% occlusion on the training and test data separately, pairwise SVM still obtains high average recognition rate with 20 times random experiments on both the USTB I and IIT Delhi I data sets. Along with the further increase in occlusion, the recognition rate begins to drop, especially on the USTB I data set. One possible explanation may be that the ear images in the IIT Delhi I data set contain more background and the random occlusion has higher probability to be applied to the background region. Moreover, it can be observed that, even with heavy occlusion (e.g. 40%), our method can still achieve competitive results of >90% average recognition rate, further demonstrating its robustness to the occlusion.

4.6 Robustness to training/testing partitions

Using the USTB I and IIT Delhi I data sets, experiments are conducted to evaluate the robustness of the proposed method to training/testing partitions. For each subject, we randomly select two ear images of each subject to constitute the training set, and use the remaining images as the test set to obtain the recognition rate. By repeating such experiment 20 times, the average recognition rate together with the standard deviation (std.) are adopted to assess the robustness. Fig. 10 shows the recognition rate of each run on the two data sets. On the USTB I data set, the average recognition rate and std. are 99.33 and 0.97%. While on the IIT Delhi I data set, the average recognition rate and std. are 99.23 and 1.01%. The results indicate that the proposed method is robust to the selection of training/test sets.

4.7 Run time

Although our method is promising in terms of identification performance, the introduction of deep feature requires more computational budget and may restrict the application of our method in real-time system. In Table 3, we evaluate the time cost of feature extraction on CPU and GPU, respectively. On a computer with 3.4 GHz Intel Xeon E3-1230 CPU and MatConvNet toolbox, our method takes 0.08–0.10 s to extract the deep feature for one image based on VGG-M, and 0.25–0.55 s based on VGG-16. Fortunately, the speed can be significantly accelerated when a GPU is deployed. With an NVIDIA GeForce GTX 1080 Ti GPU, the feature extraction time decreases to 0.003–0.005 s for VGG-M, and 0.007–0.009 s for VGG-16, and undoubtedly can be performed in real time. Furthermore, even for embedded biometric system, one can exploit FPGA-based implementation [49] to meet the real-time requirement.

4.8 Comparison with the state-of-the-arts

In this subsection, we compare the proposed method with several state-of-the-art ear recognition approaches. The results of the proposed method are obtained by fusing the deep features from fc6 and fc7 layers. The positive pairs composed are repeated twice for data augmentation and the ratio between the positive and negative pairs is set to 1:3.

**Scenario 1**: In this scenario, one ear image of each subject is chosen for training and the remaining ones are used for testing. Since the subjects in these four databases have at least three images, most competing methods [24, 50] conduct three experiments by, respectively, using the first, the second, or the third images per subject for training. Then, the average recognition rate is given in [24, 50]. We also adopt such protocol and report the average recognition rates in Table 4. In addition, due to the requirement of the pairwise data for pairwise SVM, we only conduct SVM and the k-nearest neighbour (k = 1) (K-NN) method by fusing the fc6 and fc7 deep features with the DCA algorithm. In the evaluation of the K-NN algorithm, we provide the rank-1, rank-2, and rank-3 recognition rates. As shown in Table 4, the performance obtained by SVM and KNN with feature fusion is favourably against other state-of-the-art methods on all data sets.
Table 4: Recognition rates (%) for scenario 1

<table>
<thead>
<tr>
<th>Methods</th>
<th>IIT Delhi I</th>
<th>IIT Delhi II</th>
<th>USTB I</th>
<th>USTB II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mu et al. [51]</td>
<td>–</td>
<td>–</td>
<td>85.0</td>
<td></td>
</tr>
<tr>
<td>Benzaoui et al. [24]</td>
<td>94.3</td>
<td>92.6</td>
<td>96.0</td>
<td>–</td>
</tr>
<tr>
<td>Benzaoui et al. [50]</td>
<td>91.3</td>
<td>90.3</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>our approach</td>
<td>SVM</td>
<td>92.4</td>
<td>93.7</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td>K-NN, rank-1</td>
<td>92.4</td>
<td>93.7</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td>K-NN, rank-2</td>
<td>94.4</td>
<td>96.4</td>
<td>94.2</td>
</tr>
<tr>
<td></td>
<td>K-NN, rank-3</td>
<td>94.8</td>
<td>98.2</td>
<td>95.8</td>
</tr>
</tbody>
</table>

The bold values indicate the best results.

Table 5: Recognition rates (%) for scenario 2

<table>
<thead>
<tr>
<th>Methods</th>
<th>IIT Delhi</th>
<th>IIT Delhi II</th>
<th>USTB I</th>
<th>USTB II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang et al. [52]</td>
<td>PCA</td>
<td>–</td>
<td>85.0</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>ICA</td>
<td>–</td>
<td>88.3</td>
<td>–</td>
</tr>
<tr>
<td>Nosrati et al. [53]</td>
<td>PCA</td>
<td>–</td>
<td>–</td>
<td>84.3</td>
</tr>
<tr>
<td></td>
<td>ICA</td>
<td>–</td>
<td>–</td>
<td>87.7</td>
</tr>
<tr>
<td>2D wavelet + PCA</td>
<td>–</td>
<td>–</td>
<td>90.5</td>
<td>–</td>
</tr>
<tr>
<td>Kumar and Wu [54]</td>
<td>shape features</td>
<td>29.1</td>
<td>30.6</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Gabor</td>
<td>90.4</td>
<td>88.4</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Log Gabor</td>
<td>96.3</td>
<td>95.8</td>
<td>–</td>
</tr>
<tr>
<td>Benzaoui et al. [24]</td>
<td>97.3</td>
<td>97.3</td>
<td>98.3</td>
<td>–</td>
</tr>
<tr>
<td>Benzaoui et al. [50]</td>
<td>96.7</td>
<td>97.3</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Ghoualmi et al. [22]</td>
<td>–</td>
<td>–</td>
<td>97.2</td>
<td>–</td>
</tr>
<tr>
<td>Tariq et al. [55]</td>
<td>95.2</td>
<td>–</td>
<td>98.3</td>
<td>–</td>
</tr>
<tr>
<td>our approach</td>
<td>SVM</td>
<td>94.4</td>
<td>96.8</td>
<td>96.7</td>
</tr>
<tr>
<td>pairwise SVM</td>
<td>99.9</td>
<td>99.8</td>
<td>99.4</td>
<td>99.6</td>
</tr>
<tr>
<td>K-NN, rank-1</td>
<td>94.4</td>
<td>97.7</td>
<td>96.7</td>
<td>87.7</td>
</tr>
<tr>
<td>K-NN, rank-2</td>
<td>94.4</td>
<td>98.6</td>
<td>96.7</td>
<td>90.3</td>
</tr>
<tr>
<td>K-NN, rank-3</td>
<td>94.4</td>
<td>99.1</td>
<td>96.7</td>
<td>90.9</td>
</tr>
</tbody>
</table>

The bold values indicate the best results.

Table 6: Recognition rates (%) for scenario 3

<table>
<thead>
<tr>
<th>Methods</th>
<th>IIT Delhi</th>
<th>USTB II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang and Mu [37]</td>
<td>PCA</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>ICA</td>
<td>92.2</td>
</tr>
<tr>
<td>Boodoo and Baichoo [29]</td>
<td>PCA</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td>LBP</td>
<td>93.0</td>
</tr>
<tr>
<td>Ghoualmi et al. [22]</td>
<td>HE</td>
<td>99.1</td>
</tr>
<tr>
<td></td>
<td>CLAHE</td>
<td>98.2</td>
</tr>
<tr>
<td></td>
<td>ABC</td>
<td>99.6</td>
</tr>
<tr>
<td>Guo and Xu [56]</td>
<td>–</td>
<td>93.2</td>
</tr>
<tr>
<td>Tariq et al. [55]</td>
<td>–</td>
<td>96.1</td>
</tr>
<tr>
<td>our approach</td>
<td>CNN + DCA</td>
<td>SVM</td>
</tr>
<tr>
<td>pairwise SVM</td>
<td>99.5</td>
<td>99.0</td>
</tr>
<tr>
<td>K-NN, rank-1</td>
<td>95.0</td>
<td>97.4</td>
</tr>
<tr>
<td>K-NN, rank-2</td>
<td>96.6</td>
<td>98.7</td>
</tr>
<tr>
<td>K-NN, rank-3</td>
<td>97.5</td>
<td>98.7</td>
</tr>
</tbody>
</table>

The bold values indicate the best results.

Scenario 2: This scenario is widely adopted to evaluate the ear recognition task. However, some competing methods only state to use about two ear images per subject for training, but do not provide more details on training/testing partitions [22, 52, 53]. In [24, 50, 54], the average recognition rates are obtained based on the three training/testing partitions by, respectively, using the first and the second (the first and the third, or the second and third) images per subject for training. For fair comparison, we adopt the same protocol in [24, 50, 54], and report the average recognition rate of the proposed method in Table 5. From Table 5, we can observe the performance gain by pairwise SVM over the other state-of-the-art methods on the four databases. The pairwise SVM can achieve the recognition rates of 99.86, 99.82, 99.44, and 99.57% on the USTB I, USTB II, IIT Delhi I, and IIT Delhi II data sets, respectively. Furthermore, compared with the standard SVM and K-NN classifier, pairwise SVM also shows its superiority with a large margin, which verifies the effectiveness of the feature fusion strategy and the pairwise SVM classifier.

We also report the DIR–FAR curves of the proposed method for scenario 2. Owing to the lack of test images, we perform 20 times random experiments as in the ablative evaluation to obtain Fig. 11. Since the source codes of the competing methods are not available, we also report the DIR–FAR curves of the proposed method for the four data sets. One can see that the DIRs by our method generally are high across the data sets. Specifically, our method achieve DIR at 1% of 98.17, 100.00, 99.59, and 99.75% on the USTB I, USTB II, IIT Delhi I, and IIT Delhi II data sets, respectively.

Scenario 3: Finally, we utilise three images per subject for training and the remaining for testing on the IIT Delhi I and USTB II databases. Comparison experiments are conducted with several state-of-the-art approaches. It should be noted that, the competing methods adopt different settings and do not provide the details on training/testing partitions. For fair comparison, we use the similar protocol for scenario 2, and perform three experiments based on the three partitions of training/test sets to obtain the average recognition rate. In each partition, three image per subject are randomly selected to constitute the training set. As presented in Table 6, the proposed method gets the average recognition rate of >99.00% on both data sets and performs favourably against the state-of-the-arts. Furthermore, we conduct 20 times random experiments to get Fig. 12, which shows the DIR–FAR curves of the proposed method on the two data sets. The proposed method achieves the DIR at 1% of 99.94 and 99.87% on the USTB II and IIT Delhi I data sets, respectively.
5 Conclusion

In this paper, we present an ear recognition method by fusing deep features from different layers for enhanced ear representation. With the detailed analyses on the features from different layers, we apply the DCA method to fuse different features. Finally, due to the lack of training samples of ear images, we propose to solve the ear identification problem by employing pairwise SVM. Experiments are performed on four public databases, USTB I, USTB II, IIT Delhi I, and IIT Delhi II. The results show that the proposed method can achieve promising recognition rates and performs favourably over the state-of-the-art ear recognition methods.

6 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant nos. 61671182, 61471146, and 51308096. The authors thank the anonymous reviewers for their valuable and reconstructive suggestions.

7 References