Multimed Tools Appl
DOI 10.1007/s11042-015-2650-1

Ensemble based extreme learning machine for cross-modality face matching

Yi Jin[1] · Jiuwen Cao[2] · Yizhi Wang[1] · Ruicong Zhi[3]

Received: 9 January 2015 / Revised: 13 April 2015 / Accepted: 20 April 2015

Abstract Extreme learning machine (ELM) is one of the most important and efficient machine learning algorithms for pattern classification due to its fast learning speed. In this paper, we propose a new ensemble based ELM approach for cross-modality face matching. Different to traditional face recognition methods, the proposed approach integrates the voting-base extreme learning machine (V-ELM) with a novel feature learning based face descriptor. Firstly, the discriminant feature learning is proposed to learn the cross-modality feature representation. Then, we used common subspace learning based method to reduce the obtained cross-modality features. Finally, Voting ELM is utilized as the classifier to improve the recognition accuracy and to speed up the feature learning process. Experiments conducted on two different heterogeneous face recognition scenarios demonstrate the effectiveness of our proposed approach.

✉ Yi Jin
yjin@bjtu.edu.cn

Jiuwen Cao
jwcao@hdu.edu.cn

Yizhi Wang
yzwang@bjtu.edu.cn

Ruicong Zhi
zhirc_research@hotmail.com

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044 China

[2] Institute of Information and Control, Hangzhou Dianzi University, Zhejiang, 310018, China

[3] China National Institute of Standardization, Beijing, 100191, China

Published online: 10 May 2015

Keywords Extreme learning machine · Neural network · Cross-modality matching · Feature learning · Canonical correlation analysis

1 Introduction

As a newly emerged biometric application, the cross-modality face matching [17] also called heterogeneous face recognition (HFR) has attracted much attention over the last decades for its wide range of usage in surveillance systems. Cross-modality face matching involves matching face images from alternate image modalities, such as infrared images to visible light images, sketches to photos, and 3D range images to 2D photographs. However, the performances of conventional face recognition algorithms decrease largely due to the appearance differences of cross-modality images. To address this issue, a number of HFR methods [4, 6, 13, 17, 26, 27] have been developed to solve the cross-modality matching problem. These methods generally fall into three categories: 1) homogenous image synthesis [6, 39, 40], 2) common subspace learning [13, 19, 20, 27, 34], and 3) modality-invariant feature extraction [12, 16, 18]. Homogenous image synthesis based methods generate pseudo-homogenous images, and thus, the cross-modal matching problem can be solved by using the existing FR algorithms. The common subspace learning based methods try to learn a coupled common subspace in which the cross-modality data points are considered to be more comparable than in their original representations. And the modality-invariant feature extraction based methods address the cross-modality FR problem by designing an effective invariant descriptor, and reducing the appearance differences in the feature representation stage.

Most of these existing methods try to solve this problem by deducing the cross-modality feature gap, and they have not considered the similarity measure between heterogeneous features. Recently, Extreme learning machines (ELM) [10, 11] with its high learning efficiency in feature classification have attracted increasing attention from worldwide researchers. ELM algorithms have good generalization performance in many real applications. However, very few work that considers both the feature representation and the similarity measure has been reported in the HFR research community. In this paper, we propose a new ELM ensemble based approach for cross-modality face matching. There are two stages in our proposed framework. In the first stage, we consider the cross-modality feature representation by a data-driven way, namely, the feature descriptor is optimally learned from the two modalities at the image pixel level. In the second stage, the voting based ELM is implemented as the classifier for the cross-modality face recognition.

The remainder of this paper is organized as follows. Section 2 is the related work and Section 3 describes the new ELM ensemble based approach. Experimental results and discussions on two different heterogeneous face databases are presented in Section 4. Section 5 draws the conclusion of this paper.

2 Related work

2.1 Cross-modality face matching

Previous work on cross-modality face matching can be grouped to three categories: 1) homogenous image synthesis, 2) common subspace learning, and 3) invariant feature

extraction. Most of these approaches can be organized into two steps, namely, the cross-modality feature representation and the follow-up classification.

The typical synthesis methods usually represent the data in either of the two modality, the synthesis data can then be compared directly in one modality. For instance, Tang and Wang proposed an eigen-transformation method that synthesized pseudo sketch images from the training photo sets [39] and a photo-sketch transformation method using a multi-scale Markov Random Fields (MRF) model [40]. Liu et al. [29] proposed to generate the sketches from photographs using a local linear embedding method. Gao et al. [6] utilized the embedded hidden Markov model (E-HMM) to learn the nonlinear relationship between a sketch and its corresponding photo. However, most of the synthesis methods are "task specific", which are usually designed for two fixed modalities and not generalized well when the task is changed.

The second category, common subspace learning methods represent the feature points by projecting them into a common discriminant subspace [19, 20, 27, 34]. Subspace learning approaches, such as Canonical Correlation Analysis (CCA) [7, 8] and Partial Least Squares (PLS) [41], have been approved as an effective tool in cross-modality tasks [23, 34, 36]. In Ref. [27], Lin and Tang proposed to solve the inter-modality problem using Common Discriminant Feature Extraction (CDFE), which formulated the learning objective by incorporating both the discriminative ability and the local consistency. Lei and Li et al. proposed the coupled spectral regression (CSR) [19] which modeled the properties of heterogeneous data separately by learning two associated projections. Later, they proposed the coupled Discriminant Analysis (CDA) [20] by incorporating the Locality Constraint in Kernel Space to improve the generalization ability. Even though these approaches have shown good performance in HFR, they ignore the intuitive appearance differences at the feature level. And if the cross-modality difference at the feature level is large, the discriminant power of the subspace learning methods will be reduced largely.

Methods in the third category try to reduce the cross-modality gap at the feature extraction stage. Many local appearance descriptors, e.g. variants of Local Binary Patterns (LBP) [1],SIFT [31] and Difference of Gaussian (DOG) filter [26], are utilized to represent the cross-modality features. Klare et al. [18] proposed to extract the SIFT and Multiscale LBP for forensic sketch and mug shot photo matching. Huang et al. [12] proposed to learn modality-invariant features (MIF) for HFR. B.F. Klare et al. proposed a kernel prototype similarities based generic framework [16] which introduces two filters and three different feature descriptors for feature extraction. Zhu et al. [48] proposed a feature representation method using three steps, namely, Log-DoG filtering, local encoding and uniform feature normalization. Li et al. [25] proposed to extract the common features from cross-modality face images and applied it onto optical face images and infrared face images matching. Yi et al. [43] proposed to use a series of local RBMs to learn the shared representation of two different modalities. However, most of these local descriptors are pre-defined in a hand-crafted way and they may not be the optimal one to extract the inter-modality variations.

2.2 Extreme learning machines (ELM)

In this subsection, we briefly review the ELM and its applications on pattern classification [9, 11]. ELM is recently proposed for efficiently training single-hidden-layer feed forward neural networks (SLFNs). And ELM performs more consistently with a much faster training speed [9]. The essence of ELM is that ELM performs classification by projecting original data to a high dimensional vector and changes the classification task into a multi-output functional regression problem [2].

With its high learning efficiency, ELM [10, 11] has attracted increasing attention on a widespread type of applications, e.g. pattern classification, object recognition, data analysis et.al. Huang et al. [11] extended ELM to Least square SVM (LS-SVM) [37] and proximal SVM (PSVM) [5], and provided a unified solution for multi-class classification. Kasun et al. [15] proposed an ELM based Auto Encoder (ELM-AE) for Big Data application. Cao et al. [3] proposed an improved ELM based method using the basic ELM and the OP-ELM and applied the algorithm for protein sequence classification. Later, researchers have proposed the ensemble based ELM or the ELM ensemble [28], which connected the ELM network in parallel and consider the average of the ELMs outputs as the final predicted result [9]. For example, Yang et al. [42] proposed a modified ensemble of extreme learning machine (ELM) based on attractive and repulsive particle swarm optimization (ARPSO) to improve the convergence performance of the ensemble system. Zhang et al. [45] conducted a robust AdaBoost.RT based ensemble ELM (RAE-ELM), which combined ELM with the novel self-adaptive AdaBoost.RT algorithm to achieve a better performance for regression problems.

Many ELM based approaches are proposed in FR tasks, such as, Zong and Huang [46] proposed a ELM based method in multi-label FR applications. Zong et.al. [47] later proposed a kernelized ELM method in FR. Mohammed et.al. [32] proposed a bidirectional 2DPCA and ELM framework by using curvelet feature. Long et.al. [30] proposed a graph regularized discriminative non-negative matrix factorization (GDNMF), where the projection matrix is learned jointly by both the graph Laplacian and supervised label information.

However, these methods can not be utilized directly for cross-modality face recognition due to the appearance difference in different modalities. Meanwhile, a single ELM can be improved to achieve better generalization performance [2, 14, 44]. In this paper, we propose a new ensemble ELM based approach, which is also a feature learning based ensemble ELM, for cross-modality face matching. The complete discriminative feature learning (CDFL) is used to extract the cross-modality facial features. The voting based extreme learning machine (V-ELM) [2] is utilized to perform the final image classification. Compared to other neural network based HFR methods, the proposed method requires less computational time and obtains better accuracy.

3 The proposed ensemble ELM based approach

In this section, we first introduce the basic formulation and the optimization of our Complete Discriminative Feature Learning (CDFL). Then, we explain how to use V-ELM for feature classification. Finally, the whole ensemble ELM based approach is presented (Fig. 1), which illustrates the whole process of our approach for cross-modality face matching.

3.1 Complete discriminative feature learning for feature representation

Given an $p \times q$ image M and $\mathscr{L}(M)$ is the filtered image of m. Suppose the discriminative image filter vectors to be ℓ and $\ell = [\ell_1, \ell_2, \cdots, \ell_k]^T$, the value of the filtered image at position k is,

$$\mathscr{L}(M)^n = \ell^T M^n \tag{1}$$

where M^n is the image patch centered at position n. Considering the LBP feature extraction process, the pixel difference vector (PDV) [21, 22] can be defined as,

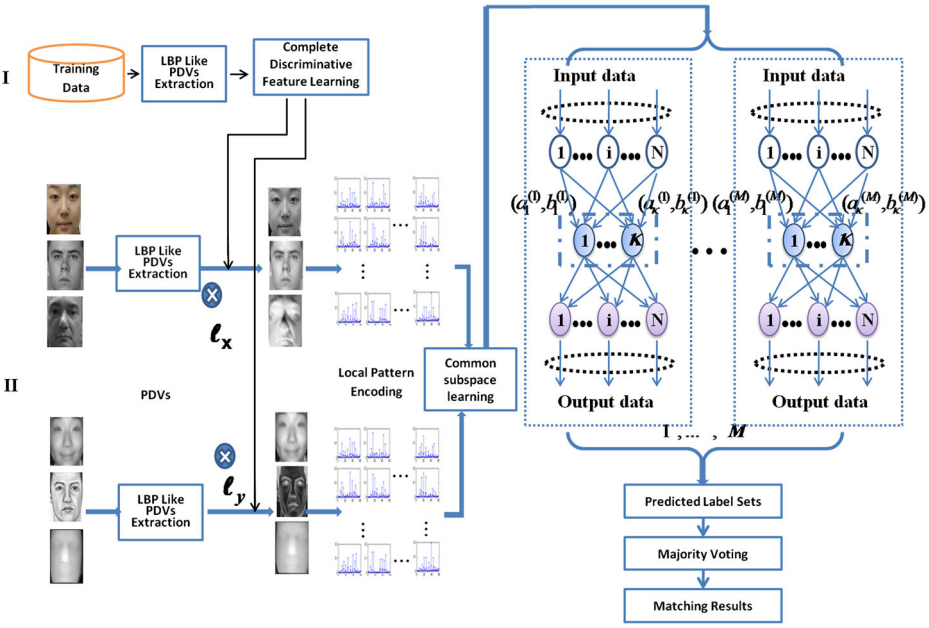

Fig. 1 The whole process of the new ensemble ELM based approach for cross-modality face matching.Part I is the feature learning phase and Part II is the face matching phase

$d\mathscr{L}(M)^n = \left[\mathscr{L}(M)^{n_1} - \mathscr{L}(M)^n, \mathscr{L}(M)^{n_2} - \mathscr{L}(M)^n, \cdots, \mathscr{L}(M)^{n_L} - \mathscr{L}(M)^n\right]^T$ where $\mathscr{L}(M)^n$ and $\mathscr{L}(M)^{n_d}$ is the pixel value of filtered image at the center n and the n_d *-th* position n_d, $n_d \in \{n_1, n_2, \cdots n_L\}$, and L is the number of neighbours. Under the linear assumption, it's quite natural to deduce that the PDV can be represented as:

$$d\mathscr{L}(M)_{ij}^n = \ell^T d(M)_{ij}^n \tag{2}$$

where $d(M)_{ij}^n$ is the *n-th* PDV of *j-th* sample from the *i-th* class, $d(M)_{ij}^n = \left[\left((M)_{ij}^{n_1} - (M)_{ij}^n\right), \left((M)_{ij}^{n_2} - (M)_{ij}^n\right), \cdots, \left((M)_{ij}^{n_L} - (M)_{ij}^n\right)\right]^T$.

The goal of complete discriminant feature learning (CDFL) is to find the optimal combined discriminative filter ℓ that can make the image PDVs of the same person similar in different modalities, so that the discriminant pixels are strengthened and the undistinguishable ones are suppressed, which makes the mapping simplified. The CDFL is defined as the following,

$$\ell_t = \sum_{t=1}^{k} U_{g_t} V_t \tag{3}$$

where k and g_t are the numbers of the discriminant filters and the *t-th* row vector of the filter graph, separately. $g = \{g_1, \cdots, g_t, \cdots, g_k\}$ contains M discriminative filters. Thus, U can be considered as the Matrix that is consisted of the discriminative image filters and V is the projection coefficients that can be treated as the weights of U.

3.1.1 Discriminative filters learning

The samples of the two sets are defined as, $M^x = \left[M_1^x, M_2^x, \cdots, M_{N_x}^x\right]$ and $M^y = \left[M_1^y, M_2^y, \cdots, M_{N_y}^y\right]$, where M^x and M^y indicate two different image modalities and N_x and N_y is the number of samples. According to (1) and (2), the discriminative filter learning aims to find an optimal image filter vector U, which can naturally split into a pair of image filter vectors U_x and U_y as $U = \left[U_x; U_y\right]$.

Given C classes of heterogenous faces, and N_i is the number of samples from i-th classes. The intra-modality and cross-modality within class and between class scatters of the filtered image are denoted as,

$$
\begin{aligned}
{G_w}^{xx} &= \sum_{i=1}^{C}\sum_{j=1}^{N_i}\left(d\mathscr{L}(M^x)_{ij} - d\mathscr{L}(\bar{M}^x)_i\right)\left(d\mathscr{L}(M^x)_{ij} - d\mathscr{L}(\bar{M}^x)_i\right)^T \\
{G_w}^{xy} &= \sum_{i=1}^{C}\sum_{j=1}^{N_i}\left(d\varphi(P^x)_{ij} - d\mathscr{L}(\bar{M}^y)_i\right)\left(d\mathscr{L}\left(P^x\right)_{ij} - d\varphi\left(\bar{M}^y\right)_i\right)^T
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
{G_b}^{xx} &= \sum_{i=1}^{C} C_i\left(d\mathscr{L}\left(\bar{M}^x\right)_i - d\mathscr{L}\left(\bar{M}^x\right)\right)\left(d\mathscr{L}\left(\bar{M}^x\right)_i - d\mathscr{L}\left(\bar{M}^x\right)\right)^T \\
{G_b}^{xy} &= \sum_{i=1}^{C} C_i\left(d\mathscr{L}\left(\bar{M}^x\right)_i - d\mathscr{L}\left(\bar{M}^y\right)\right)\left(d\mathscr{L}\left(\bar{M}^x\right)_i - d\mathscr{L}\left(\bar{M}^y\right)\right)^T
\end{aligned} \tag{5}
$$

${G_w}^{yy}$ and ${G_w}^{yx}$ are similar to ${G_w}^{xx}$ and ${G_w}^{xy}$ according to (4). And ${G_b}^{yy}$ and ${G_b}^{yx}$ can be defined similarly as (5). $d\mathscr{L}(M^x)_{ij}$ and $d\mathscr{L}(M^y)_{ij}$ are the Pixel Difference Matrixes (PDM) of the j-th sample pair from the i-th class, and $\mathscr{L}(\bar{M}^x)_i$, $\mathscr{L}(\bar{M}^y)_i$ are the mean matrixes of the PDVs on the filtered image from the i-th class, $\mathscr{L}(\bar{M}^x)$ and $\mathscr{L}(\bar{M}^y)$ are the total mean vectors of the PDVs of the same sample set.

The within class scatter and between class scatter of the filter images can be defined as,

$$
\boldsymbol{G_w} = \begin{bmatrix} G_w^{xx} & G_w^{xy} \\ G_w^{yx} & G_w^{yy} \end{bmatrix}, \boldsymbol{G_b} = \begin{bmatrix} G_b^{xx} & G_b^{xy} \\ G_b^{yx} & G_b^{yy} \end{bmatrix} \tag{6}
$$

According to (1), we get:

$$
\boldsymbol{G_w} = U^T\tilde{\boldsymbol{G}}_{\boldsymbol{w}}U, \boldsymbol{G_b} = U^T\tilde{\boldsymbol{G}}_{\boldsymbol{b}}U \tag{7}
$$

And $\tilde{\boldsymbol{G}}_{\boldsymbol{w}}$ and $\tilde{\boldsymbol{G}}_{\boldsymbol{b}}$ is defined as $\tilde{\boldsymbol{G}}_{\boldsymbol{w}} = \begin{bmatrix} \tilde{G}_w^{xx} & \tilde{G}_w^{xy} \\ \tilde{G}_w^{yx} & \tilde{G}_w^{yy} \end{bmatrix}$, $\tilde{\boldsymbol{G}}_{\boldsymbol{b}} = \begin{bmatrix} \tilde{G}_b^{xx} & \tilde{G}_b^{xy} \\ \tilde{G}_b^{yx} & \tilde{G}_b^{yy} \end{bmatrix}$. And (4) and (5) are changed into:

$$
\begin{aligned}
{G_w}^{xx} &= U_x^T\sum_{i=1}^{C}\sum_{j=1}^{N_i}(d(M^x)_{ij} - d(\bar{M}^x)_i)(d(M^x)_{ij} - d(\bar{M}^x)_i)^T U_x = U_x^T\tilde{G}_w^{xx}U_x \\
{G_w}^{xy} &= U_x^T\sum_{i=1}^{C}\sum_{j=1}^{N_i}(d(M^x)_{ij} - d(\bar{M}^y)_i)(d(M^x)_{ij} - d(\bar{M}^y)_i)^T U_y = U_x^T\tilde{G}_w^{xy}U_y
\end{aligned} \tag{8}
$$

and

$$G_b{}^{xx} = U_x^T \sum_{i=1}^{C} C_i (d(\bar{M}^x)_i - d(\bar{M}^x))(d(\bar{M}^x)_i - d(m^x))^T U_x = U_x^T \tilde{G}_b^{xx} U_x$$
$$G_b{}^{xy} = U_x^T \sum_{i=1}^{C} C_i (d(\bar{M}^x)_i - d(\bar{M}^y))(d(\bar{M}^x)_i - d(\bar{M}^y))^T u_y = U_x^T \tilde{G}_b^{xy} U_y \quad (9)$$

Finally, the problem of discriminative filters can be achieved by solving the generalized eigenvalue problem of $\tilde{\boldsymbol{G}}_{\boldsymbol{b}} U_t = \lambda_1 \tilde{\boldsymbol{G}}_{\boldsymbol{w}} U_t$.

3.1.2 Enhanced weight learning

As it is in (3), the CDFL can be treated as a weighted sum of all these discriminative filters U using a linear projection from M-dimensional subspace to 1-dimensional vector. And the weight vector V can be considered as the projection coefficients of the linear projection. While the discriminative image filter learning aims at learning the discriminant pixels in each small image patch, it fails to notice the discrimination in both the intra-personal and inter-personal PDV pairs. Therefore, the aim of weighting the discriminative filters is to make the similarities of PDVs from two classes (the intra-personal and inter-personal PDV pairs) more discriminable, and the weight learning problem of the CDFL can then be transformed to a two class linear projection.

Give M groups of Z, Z is the similarities of PDVs pairs. $\mathrm{Z} = [\mathrm{Z}_1; \mathrm{Z}_2; \cdots; \mathrm{Z}_M]^T$, and $\mathrm{Z}_\gamma = [\mathrm{Z}^{\mathrm{intra}}{}_\gamma, \mathrm{Z}^{\mathrm{inter}}{}_\gamma]$, $\gamma \in \{1, 2, \cdots, m\}$, the samples of the two classes (the intra-personal and inter-personal PDV pairs) are defined as,

$$\begin{cases} \mathrm{Z}^{\mathrm{intra}}{}_\gamma = [\theta_1^{\mathrm{intra}}, \theta_2^{\mathrm{intra}}, \cdots, \theta_{N_{\mathrm{intra}}}^{\mathrm{intra}}] \\ \mathrm{Z}^{\mathrm{inter}}{}_\gamma = [\theta_1^{\mathrm{inter}}, \theta_2^{\mathrm{inter}}, \cdots, \theta_{N_{\mathrm{intra}}}^{\mathrm{inter}}] \end{cases} \quad (10)$$

where θ^{intra} and θ^{inter} indicate the similarities of the intra-personal and inter-personal PDV pairs, which are defined as,

$$\begin{cases} \theta^{\mathrm{intra}} = \left\| d(M)_{ij}^n - d(M)_{\mu\upsilon}^n \right\|, \text{where } i = \mu \text{ and } j \neq \upsilon \\ \theta^{\mathrm{inter}} = \left\| d(M)_{ij}^n - d(M)_{\mu\upsilon}^n \right\|, \text{where } i \neq \mu \end{cases} \quad (11)$$

and N_{intra} and N_{inter} is the number of PDV pairs from the two classes. To address this problem, we utilize the Fisher Linear Discriminant Analysis (FLDA) to get the optimal linear projection. The between class scatter matrix and within class scatter matrix of P are defined as:

$$G_b{}' = \left(\bar{\mathrm{Z}}^{\mathrm{intra}} - \bar{\mathrm{Z}}^{\mathrm{inter}}\right)\left(\bar{\mathrm{Z}}^{\mathrm{intra}} - \bar{\mathrm{Z}}^{\mathrm{inter}}\right)^T$$
$$G_w{}' = \sum_{n_i=1}^{N_{\mathrm{intra}}} \left(\mathrm{Z}_{n_i} - \bar{\mathrm{Z}}^{\mathrm{intra}}\right)\left(\mathrm{Z}_{n_i} - \bar{\mathrm{Z}}^{\mathrm{intra}}\right)^T + \sum_{n_j=1}^{N_{\mathrm{inter}}} \left(\mathrm{Z}_{n_j} - \bar{\mathrm{Z}}^{\mathrm{intra}}\right)\left(\mathrm{Z}_{n_j} - \bar{\mathrm{Z}}^{\mathrm{intra}}\right)^T \quad (12)$$

where $\bar{\mathrm{Z}}^{\mathrm{intra}}$ and $\bar{\mathrm{Z}}^{\mathrm{inter}}$ is the mean of the two classes. Thus, the projection coefficients V, which is also the optimal weights for discriminating the PDV pairs, can be obtained by solving the generalized eigenvalue problem of $G_b{}'V = U_2 G_w{}'V$.

3.2 V-ELM for classification

To tackle the issue of cross-modality face matching and improve the general performance of ELM, a voting based ELM [2] is utilized in our model for feature classification. In V-ELM, multiple indenpent ELMs are firstly trained with the a fixed number of hidden nodes and the same activation function. The learning parameters of each ELM are randomly initialized independently. Then, the predicted label is determined by a majority voting method. The V-ELM classifier utilized in the proposed recognition approach can be described as follows.

Assuming that the available training feature dataset is $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^{\mathbf{N}}$, where $\boldsymbol{x}_i$, t_i, and $\mathbf{N}$ represent the feature vector of the i-th face image, its corresponding category index and the number of images, respectively, the SLFN with κ nodes in the hidden layer can be expressed as

$$\boldsymbol{o}_i = \sum_{j=1}^{\kappa} \theta_j g(\boldsymbol{a}_j, \boldsymbol{b}_j, \boldsymbol{x}_i), i = 1, 2, \cdots, \mathbf{N} \tag{13}$$

where $\boldsymbol{o}_i$ is the output obtained by the SLFN associated with the i-th input protein sequence, $\boldsymbol{a}_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$ $(j = 1, 2, \cdots, \kappa)$ are parameters of the jth hidden node, respectively. The variable $\boldsymbol{\theta}_j \in \mathbb{R}^m$ is the link connecting the jth hidden node to the output layer and $g(\cdot)$ is the hidden node activation function. With all training samples, (13) can be expressed in the compact form as

$$\mathbf{O} = \boldsymbol{H}\boldsymbol{\theta} \tag{14}$$

where $\theta = (\theta_1, \theta_2, \cdots, \theta_\kappa)$ and $\mathbf{O}$ are the output weight matrix and the network outputs, respectively. The variable $\boldsymbol{H}$ denotes the hidden layer output matrix with the entry $\boldsymbol{H}_{ij} = g(\boldsymbol{a}_j, b_j, \boldsymbol{x}_i)$.

To perform multi-classes classification, the ELM classifier generally utilizes the One-Against-All (OAA) method to transform the classification application to a multi-output model regression problem. That is, for a C-categories classification application, the output label t_i of the face image feature $\boldsymbol{x}_i$ is encoded to a $\mathcal{C}$-dimensional vector $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \cdots, t_{i\mathcal{C}})^T$ with $t_{i\mathbf{c}} \in \{1, -1\}$ $(\mathbf{c} = 1, 2, \cdots, \mathcal{C})$. If the category index of the face image $\boldsymbol{x}_i$ is $\mathbf{c}$, then $t_{i\mathbf{c}}$ is set to be 1 while the rest entries in $\boldsymbol{t}_i$ are set to be -1. Hence, the objective of training phase for the SLFN in (13) becomes finding the best network parameters set $\Delta = \left\{\left(\boldsymbol{a}_j, b_j, \boldsymbol{\theta}_j\right)\right\}_{j=1,\ldots,\kappa}$ such that the following error cost function is minimized

$$\min_{\Delta} E = \min_{\Delta} \|\mathbf{O} - \boldsymbol{T}\| \tag{15}$$

where $\boldsymbol{T} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_{\mathbf{N}})$ is the target output matrix.

ELM theory claims that random hidden node parameters can be utilized for SLFNs and the hidden node parameters may not need to be tuned. In such case, the system (14) becomes a linear model and the network parameter matrix can be analytically solved by using the least-square method. That is,

$$\boldsymbol{\theta} = \boldsymbol{H}^{\dagger}\boldsymbol{T} \tag{16}$$

where $\boldsymbol{H}^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix $\boldsymbol{H}$ given by [35]. The universal approximation property of the ELM algorithm is also presented in [11].

Suppose that $\boldsymbol{M}$ independent networks trained with the ELM algorithm are used in V-ELM. Then, for each testing sample x_{test}, $\boldsymbol{M}$ prediction results can be obtained based on these independent ELMs. A corresponding vector $\pi_{\boldsymbol{M},x_{test}} \in \mathbf{R}^C$ with dimension equal to the number of class labels is used to store all these $\boldsymbol{M}$ results of x_{test}, where if the class

label predicted by the $\boldsymbol{M}^{th}$ ($m \in [1, 2, \cdots, \boldsymbol{M}]$) ELM is τ, the value of the corresponding entry τ in the vector $\pi_{\boldsymbol{M},x_{test}}$ is increased by one, as the following:

$$\pi_{\boldsymbol{M},x_{test}}(\tau) = \pi_{\boldsymbol{M},x_{test}}(\tau) + 1$$

After all these M results are achieved and assigned to $\pi_{\boldsymbol{M},x_{test}}$, the final class result of x_{test} is then achieved by conducting a majority voting:

$$C_{test} = \arg \max_{\tau \in [1,\cdots,C]} \{\pi_{\mathbf{M}}, x_{test}(\tau)\} \tag{17}$$

3.3 The ensemble ELM based approach for cross-modality face matching

The whole procedure of ensemble ELM approach can be divided into two parts, 1) the feature learning by CDFL and 2) the V-ELM based cross-modality face matching. Figure 1 illustrates the whole process of the ensemble ELM based approach for cross-modality face matching.

In the feature learning phase, face images are firstly divided into small patches and a LBP-like feature extraction is used to get the pixel difference vector (PDV). The CDFL image filters are then learned from the pixel difference matrix (PDM). In the matching phase, we firstly get the new image pattern by the learned filter matrix. Then, histogram-based local features are extracted to boost the local features. Thirdly, CCA [7] is utilized to map the data onto a common subspace and reduce the feature dimensions. Finally, V-ELM is applied to get the final matching results.

4 Experiments

In this section, we compare our ensemble ELM based approach with some state-of-the-art HFR methods, such as PLS [34], CDFE [27], CCA [7], Discriminant Image Filter Learning (DIFL) [22] and some hand-crafted feature extraction methods, e.g. LBP and LTP. Here, two different heterogeneous face recognition applications, which are VIS to NIR and 2D Vs. 3D, are utilized to evaluated the performance of our proposed method. The following describes the details of the experimental setups and the results.

4.1 Heterogeneous face bimetrics (HFB): VIS vs. NIR

The HFB database[24] is used to evaluate Visual (VIS) image vs. Near Infrared (NIR) image heterogeneous scenarios. In this experiment, a released database Ver.1 which contains images from 100 subjects, with 4 NIR and 4 VIS images per subject, is utilized. All the images are scaled, transformed and cropped to 128×128 size according to the eye position. Some of the cropped HFB example images are shown in Fig.2. In this experiments, VIS images of each person are utilized as the gallery set and the corresponding NIR images are utilized as the probe set. And the filtering window s to be $s = 3$ with which the neighbours $N_i = 8$ participates in the image filtering.

In the first experiment, the frontal 80 persons are utilized for training, and the rest 20 persons are utilized for testing. We compare the recognition performance of the proposed Ensemble ELM with several popular HFR methods, and meanwhile, we compare the new proposed method with invariant feature descriptos, such as LBP, LTP and DIFL [22]. The hidden nodes of ELM and our ensemble ELM (with Sigmoid function) are choosen to be

Fig. 2 Some of the cropped examples from HFB database

1000. Figure 3 shows the recognition rate varies with the first 100 dimensions of the several HFR methods.

As it is shown, Fig. 3a is the comparison with other HFR algorithms while Fig. 3b is the comparison with different classification methods. We can see in Fig. 3a that the proposed ensemble ELM based method significantly outperforms the other methods, and most of the compared methods get their highest recognition rate at a average range of dimension 50 to 80 except CCA which gets its highest recognition rata around dimension 30.

In particular, Table 1 gives the detailed recognition performance of different methods for the HFB database. The results from Table I indicate that ensemble ELM based approach is effective in improving cross-modality face recognition performance in general. The rank-1 recognition rate obtained by the new proposed approach is 83.8 %, which is 1.3 % and 2.5 % higher than the ELM and NN based approach.

4.2 Face recognition grand challenge (FRGC): 2D photos vs. 3d range images

The last experiments are conducted on the FRGC [33] 2D vs. 3D face database. In this experiment, FRGC v2, which contains 4007 2D and 3D face image pairs of 466 persons,

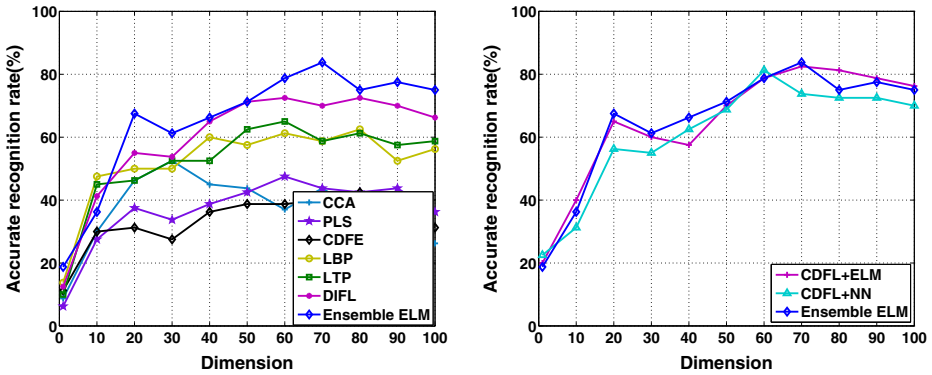

Fig. 3 Performance comparison on the HFB database, where **a** shows the recognition comparison of Ensemble ELM with different HFR methods and **b** shows the comparison of different classifiers

Table 1 Performance Comparison on Three Heterogeneous Scenarios in terms of Rank-1 Rec. Rate (in(%))

Method \ Dataset	HFB	FRGC
PLS [34]	47.5	80.1
CDFE [27]	42.5	75.7
CCA [7]	52.5	76.4
LBP [1]	62.5	83.4
LTP [38]	65.0	86.0
DIFL [22]	72.5	90.1
CDFL+NN	81.3	91.8
CDFL+ELM	82.5	92.6
Ensemble ELM	**83.8**	**95.7**

Bold entries emphasize the performance of our method, and they emphasize that our method achieves better recognition rate than the other compared methods

is utilized to evaluate the performance of the proposed method. This database consists of frontal views, expressions and et.al., but none of them is wearing glasses. All these images are cropped in the same way to 100 × 100 size according to the eye position. The cropped examples of FRGC database are shown in Fig. 4.

To evaluate the robustness of our method against expression variations, 285 subjects with more than 6 samples are picked out, and 5 samples of each person are selected as training set and the rest for testing. The hidden notes of the ELM based methods are chosen to be 500. And we repeat random selection 10 times to get an average rank-1 recognition rate. Fig. 5 shows the experimental results of the ensemble ELM based approach comparing with some popular HFR methods, such as PLS [34], CDFE [27], CCA [7]. As we can see in Fig. 5a and Fig. 5b, our ensemble ELM based method constantly outperforms the other compared HFR methods. The rank-1 recognition rate obtained by the new proposed approach is 95.7 %, which is 3.1 % and 3.9 % higher than the ELM and NN based approach. Detailed recognition results are displayed in Table 1.

4.3 Results and discussion

The recognition results of different cross-modal FR algorithms on the two databases are given in Table 1, from which we can observe that the Ensemble ELM method consistently

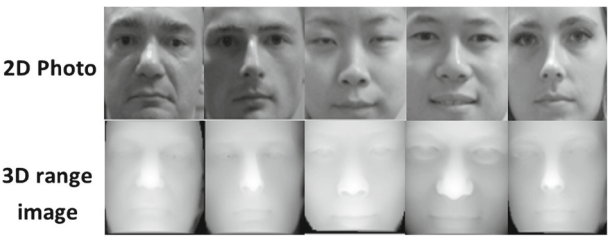

Fig. 4 The cropped examples from CUFSF database. The first row contains the examples of digital photo and the second row is the corresponding 3D range image

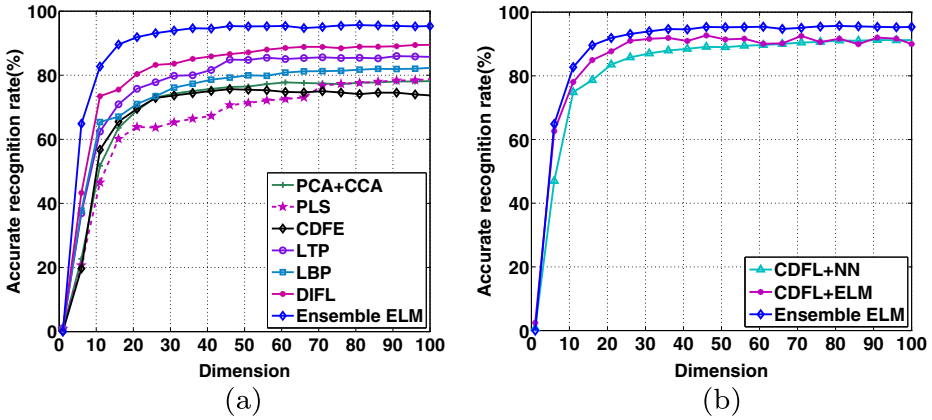

Fig. 5 Performance comparison on the FRGC database, where **a** shows the recognition comparison of Ensemble ELM with different HFR methods and **b** shows the comparison of different classification methods

outperforms the other compared HFR methods. From Table 1, we see that the rank-1 recognition rate on HFB and FRGC databases are 83.8 % and 95.7 %, respectively.

We have the following two observations from the above comparisons:

1) The proposed method provides an Ensemble ELM based approach for cross-modality image matching problems. The proposed approach solves the cross-modality FR problem from these two ways. Firstly, the cross-modality appearance differences are reduced by learning a new feature descriptor, and the cross-modality features are represented in a more discriminant way. Secondly, the recognition performance is boosted by the Ensemble ELM, which achieve better classification accuracy.
2) The proposed method consistently outperforms all other methods on two cross-modality applications. The reason lies the Ensemble ELM approach exploits the most discriminant features by using the complete discriminative feature learning. Meanwhile, the ELM are utilized in parallel and the final classification is obtained by the voting strategy.

5 Conclusion

Extreme learning machine (ELM) have good generalization performance in many real classification applications. In this paper, we have proposed a new ensemble based ELM approach for cross-modality face matching. The proposed approach exploits the combination of feature learning based face descriptors and the voting-base extreme learning machine (V-ELM). In this new approach, the cross-modality feature differences is first reduced at the image pixel level in a data-driven way, then, Voting based ELM, which has adopted multiple independent ELM training instead of a single ELM training, is utilized to achieve the cross-modality face matching result. Experiments on three different HFR applications show the effectiveness and generalization of the proposed new method.

Acknowledgments This work was supported by the fundamental research funds for the central universities (K14JB00230) and the National Natural Science Foundation of China (No. 61403024, 31201358, 61100141).

References

1. Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns:application to face recognition. IEEE Trans Pattern Anal Mach Intell 11(12):2037–2041
2. Cao J, Lin Z, Huang G-B, Liu N (2012) Voting based extreme learning machine. Inf Sci 185(1):66–77
3. Cao J, Xiong L (2014) Protein sequence classification with improved extreme learning machine algorithms. BioMed Research International, p 2014
4. Chen J, Yi D, Yang J, Zhao G, Li S, Pietikainen M. (2009) Learning mapping for face synthesis from near infrared to visual light images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 156–163
5. Fung G, Mangasarian OL (2001) Proximal support vector machine classifiers. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp 77–86
6. Gao X, Zhong J, Li J, Tian C (2005) Face sketch synthesis algorithm based on e-hmm and selective ensemble. IEEE Trans. Circuits Syst. Video Technol. 4:487–496
7. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning method. Neural Comput 16:2639–2664
8. Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377
9. Huang G-B, Wang DH, Lan Y (2011) Extreme learning machines: a survey. Int J Mach Learn Cybern 2(2):107–122
10. Huang G-B, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern Part B Cybern 42(2):513–529
11. Huang G-B, Zhu Q-Y, Siew C-K (2006) Extreme learning machine theory and applications. Neurocomputing 70(1):489–501
12. Huang LK, Lu JW, Tan Y-P (2012) Learning modality-invariant features for heterogeneous face recognition. In: Proceedings of IEEE International Conference on Pattern Recognition, pp 1683–1686
13. Huang XS, Lei Z, Fan MY, Wang X, Li SZ (2013) Regularized discriminative spectral regression method for heterogeneous face matching. IEEE Trans Image Process 22(1):353–362
14. HuiJuan L, An C, Zheng E, Yi L (2014) Dissimilarity based ensemble of extreme learning machine for gene expression data classification. Neurocomputing 128(0):22–30
15. Kasun LLC, Zhou H, Huang G-B, Vong CM (2013) Representational learning with extreme learning machine for big data. IEEE Intelligent Systems
16. Klare BF, Anil KJ (2013) Heterogeneous face recognition using kernel prototype similarities. IEEE Trans Pattern Anal Mach Intell 6:1410–1422
17. Klare BF, Jain AK (2013) Heterogeneous face recognition using kernel prototype similarities. IEEE Trans Pattern Anal Mach Intell 35(6):1410–1422
18. Klare BF, Li Z, Jain AK (2011) Matching forensic sketches to mug shot photos. IEEE Trans Pattern Anal Mach Intell 33(3):639–646
19. Lei Z, Li SZ (2009) Coupled spectral regressoin for matching heterogeneous faces. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1123–1128
20. Lei Z, Liao SC, Jain AK, Li SZ (2012) Coupled discriminant analysis for heterogeneous face recognition. IEEE Trans Inf Forensics Secur 7(6):1707–1716
21. Lei Z, Pietikainen M, Li SZ (2014) Learning discriminant face descriptor. IEEE Trans Pattern Anal Mach Intell 36(2):289–302
22. Lei Z, Yi D, Li SZ (2012) Discriminant image filter learning for face recognition with local binary pattern like representation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 2512–2517

23. Li A, Shan S, Chen X, Gao W (2011) Face recognition based on non-corresponding region matching. In: Proceedings of IEEE International Conference on Computer Vision, pp 1060–1067
24. Li SZ, Lei Z, Ao M (2009) The hfb face database for heterogeneous face biometrics research. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp 1–8
25. Li Z, Gong D, Qiao Y, Tao D (2014) Common feature discriminant analysis for matching infrared face images to optical face images. IEEE Trans Image Process 23(6):2436–2445
26. Liao S, Yi D, Lei Z, Qin R, Li S (2009) Heterogeneous face recognition from local structures of normalized appearance. In: Proceedings of International Conference on Biometrics, pp 209–218
27. Lin D, Tang X (2006) Inter-modality face recognition. In: Proceedings of the European Conference on Computer Vision, pp 13–26
28. Liu N, Wang H (2010) Ensemble based extreme learning machine. IEEE Signal Process Lett 17(8):754–757
29. Liu Q, Tang X, Jin H, Lu H, Ma S (2005) A nonlinear approach for face sketch synthesis and recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1005–1010
30. Long X, Hongtao L, Peng Y, Li W (2014) Graph regularized discriminative non-negative matrix factorization for face recognition. Multimedia Tools and Applications 72(3):2679–2699
31. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
32. Mohammed AA, Minhas R, Jonathan Wu QM, Sid-Ahmed MA (2011) Human face recognition based on multidimensional pca and extreme learning machine. Pattern Recogn 44(10–11):2588–2597
33. Phillips PJ, Flynn P, Scruggs T, Bowyer KW, Chang J, Hoffman K, Marques J, Min J, Worek W (2005) Overview of the face recognition grand challenge. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 1, pp 947–954
34. Sharma A, Jacobs DW (2011) Bypassing synthesis,Pls for face recognition with pose, low-resolution and sketch. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 593–600
35. Serre D (2002) Matrices: theory and applications. Springer
36. Sun Q, Zeng S, Liu Y, Heng PA, Xia DS (2005) A new method of feature fusion and its application in image recognition. Pattern Recogn 38(12):2437–2448
37. Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300
38. Tan X, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans Image Process 19(6):1635–1650
39. Tang X, Wang X (2004) Face sketch recognition. IEEE Trans Circuits Syst Video Technol 1:50–57
40. Wang X, Tang X (2009) Face photo-sketch synthesis and recognition. IEEE Trans Pattern Anal Mach Intell 11:1955–1967
41. Wold H (1975) Quantitative sociology: International perspectives on mathematical and statistical modeling (quantitative studies in social relations), vol 16. Academic press edn. Academic Press, London, pp 307–357
42. Yang D, Han F (2014) An improved ensemble of extreme learning machine based on attractive and repulsive particle swarm optimization. In: Intelligent Computing Theory, volume 8588 of Lecture Notes in Computer Science, pp 213–220
43. Yi D, Lei Z, Liao S, Li SZ (2014) Shared representation learning for heterogeneous face recognition. arXiv:1406.1247
44. Zhai J-h, Hong-yu X, Wang X-z (2012) Dynamic ensemble extreme learning machine based on sample entropy. Soft Comput 16(9):1493–1502
45. Zhang P, Zhixin Y (2015) Ensemble extreme learning machine based on a new self-adaptive adaboost.rt, In: Proceedings of ELM-2014 Volume 1, of Proceedings in Adaptation, Learning and Optimization, vol 3, pp 237–244
46. Zong W, Huang G-B (2011) Face recognition based on extreme learning machine. Neurocomputing 74(16):2541–2551

47. Zong W, Zhou H, Huang G-B, Lin Z (2011) Face recognition based on kernelized extreme learning machine. In: Autonomous and Intelligent Systems, of Lecture Notes in Computer Science, vol 6752, pp 263–272
48. Zhu J-Y, Zheng W-S, Lai J-H, Li SZ (2014) Matching nir face to vis face using transduction. IEEE Trans Inf Forensics Secur 9(3):501–514

Yi Jin (SM'06-M'13), received the Ph.D. degree in Signal and Information Processing from the Institute of Information Science, Beijing Jiaotong University, Beijing, P.R. China, in 2010. She is currently an Assistant Professor in the School of Computer Science and Information Technology, Beijing Jiaotong University.
She has been a visiting scholar in School of Electrical and Electronic Engineering, Nanyang Technological University of Singapore (2013-2014). She has served as the guest editor for special issues in Mathematical Problems in Engineering. Her research interests include computer vision, pattern recognition, image processing and machine learning.

Jiuwen Cao received the B. Sci. and M. Sci. in School of Applied Mathematics, University of Electronic Science and Technology of China (UESTC) in 2005 and 2008, respectively, and Ph.D degree in School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, in 2013. From Sep. 2012 to Dec. 2013, he was a research fellow in NTU. Currently, he is an Associate Professor of Hangzhou Dianzi University, China. His research interests include extreme learning machine, machine learning, neural networks, system analysis and control, and array signal processing. He has served as the guest editor for 2 special issues in Mathematical Problems in Engineering and the Program Chair for the 5th International Conference on Extreme Learning Machine, Singapore, 2014. He is now serving as a track co-chair of IEEE TENCON 2015.

Yizhi Wang received the B.S. degree from Harbin Institute of Technology, Harbin, P.R. China in 1978. She is currently a full professor in the School of Computer Science and Information Technology, Beijing Jiaotong University.

She has authored and co-authored 5 books and more than 40 technical papers in the application technology of computer science. She won the 5th national award for distinguished teachers in 2009 and was the winner of National Baogang Award for prominent teachers in 2009. Her main research interests include computer application technology, artificial intelligence, Internet applications and computer educational technology.

Ruicong Zhi received the Ph.D degree in Signal and Information Processing from Beijing Jiaotong University, Beijing, China, in 2010, the B. S. degree in biomedical engineering from Beijing Jiaotong University, Beijing, China, in 2005. From 2008 to 2009, she visited the Sound and Image Processing Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, as a joint Ph.D student. She is currently an associate researcher in China National Institute of Standardization, Beijing, China. Her research interests included sensory metrology, pattern recognition, and consumer test.