Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition

Ran He, *Senior Member, IEEE,* Xiang Wu, Zhenan Sun*, *Member, IEEE,* and Tieniu Tan, *Fellow, IEEE*

Abstract—Heterogeneous face recognition (HFR) aims at matching facial images acquired from different sensing modalities with mission-critical applications in forensics, security and commercial sectors. However, HFR presents more challenging issues than traditional face recognition because of the large intra-class variation among heterogeneous face images and the limited availability of training samples of cross-modality face image pairs. This paper proposes the novel Wasserstein convolutional neural network (WCNN) approach for learning invariant features between near-infrared (NIR) and visual (VIS) face images (i.e., NIR-VIS face recognition). The low-level layers of the WCNN are trained with widely available face images in the VIS spectrum, and the high-level layer is divided into three parts: the NIR layer, the VIS layer and the NIR-VIS shared layer. The first two layers aim at learning modality-specific features, and the NIR-VIS shared layer is designed to learn a modality-invariant feature subspace. The Wasserstein distance is introduced into the NIR-VIS shared layer to measure the dissimilarity between heterogeneous feature distributions. W-CNN learning is performed to minimize the Wasserstein distance between the NIR distribution and the VIS distribution for invariant deep feature representations of heterogeneous face images. To avoid the over-fitting problem on small-scale heterogeneous face data, a correlation prior is introduced on the fully-connected WCNN layers to reduce the size of the parameter space. This prior is implemented by a low-rank constraint in an end-to-end network. The joint formulation leads to an alternating minimization for deep feature representation at the training stage and an efficient computation for heterogeneous data at the testing stage. Extensive experiments using three challenging NIR-VIS face recognition databases demonstrate the superiority of the WCNN method over state-of-the-art methods.

Index Terms—Heterogeneous face recognition, VIS-NIR face matching, feature representation.

1 Introduction

Ubiquitous face sensors not only facilitate the wide application of face recognition but also generate various heterogeneous sets of facial images [1] [2]. Matching faces across different sensing modalities raises the problem of heterogeneous face recognition (HFR) or cross-modality face recognition. Due to significant difference in sensing processes, heterogeneous images of the same subject have large appearance variations, which has distinguished HFR from regular visual (VIS) face recognition [3]. Over the last decade, HFR has become increasingly important in many practical security applications and has attracted substantial attention in the computer vision community. Impressive progress has been made in associated research areas, such as near-infrared (NIR) vs. VIS [4], sketch vs. VIS [5], 2D vs. 3D [6], and different resolutions [7] and poses [8].

Because the NIR imaging technique provides an efficient and straightforward solution for improving face recognition performance in extreme lighting conditions, it has been considered one of the most prominent alternative sensing modalities in HFR [9]. Moreover, NIR imaging has been proved to be less sensitive to visible light illumination variations [10]. Thus this method is applicable for face recognition at a distance or even at night, and has been widely used in face identification and authorization applications, such as security surveillance and E-passports. However, most face galleries only consist of VIS images due to the mass deployment of VIS sensors, whereas probe images are often captured by NIR modalities. Therefore, the demand for robust matching between NIR and VIS face images, which is also known as the NIR-VIS HFR problem, has increased sharply and attracted considerable attention.

Substantial research efforts have been directed toward improving NIR-VIS HFR performance [2] [11]. Traditional NIR-VIS methods generally involve image synthesis, subspace learning and invariant feature extraction [12] [2]. These methods often utilize several processing steps to achieve satisfactory accuracy. Recently, inspired by the successful application of convolutional neural networks (CNNs) to VIS face recognition [13] [14] [15], several deep models [16] [17] [9] have attempted to transfer the knowledge learned on a large scale VIS face database to the NIR modality. These methods train a basic CNN network on the public CASIA NIR-VIS 2.0 database [4] and make the basic network adaptable to both NIR and VIS modalities. Experimental results suggest that deep models have the potential to outperform traditional NIR-VIS methods.

However, NIR-VIS HFR remains a challenging problem for deep models and is largely unsolved, which is mainly because of the following two reasons: **1) The gap between the sensing patterns of the VIS and NIR modalities.** Because NIR and VIS images are captured using different sensing modalities, they have large differences in feature representation. Moreover, VIS data lack the representative spectral information of NIR images; therefore, deep models trained on VIS data fail to provide satisfactory results [16] [17] [9]. The debate on the optimal measure of the difference and the approach for closing the gap between VIS and

• *Zhenan Sun is the corresponding author. R. He, X. Wu, Z. Sun and T. Tan are with National Laboratory of Pattern Recognition, CASIA, Center for Research on Intelligent Perception and Computing, CASIA, Center for Excellence in Brain Science and Intelligence Technology, CAS and University of Chinese Academy of Sciences, Beijing, China,100190. E-mail: {rhe, znsun, tnt}@nlpr.ia.ac.cn, xiang.wu@cripac.ia.ac.cn

NIR modalities remains active. Thus exploring modality-invariant representations of both NIR and VIS face images via large-scale VIS face data is difficult. **2) Over-fitting on small-scale training sets.** With the rapid development of the Internet, large collections of VIS face images can be gathered efficiently. However, VIS face images paired with NIR layouts are not widely available online, which increases the cost of obtaining paired VIS and NIR images on a large scale. Most existing HFR databases are of a small-scale (fewer than 10,000 samples) with large feature dimensions (at least 100×100 pixels). Consequently, deep models will likely over-fit the training set during feature learning [16] [9]. Determining the optimal method for fitting deep models to small-scale NIR-VIS datasets remains a central problem.

In this paper, the two aforementioned problems are tackled by a novel Wasserstein CNN (WCNN) architecture. The WCNN employs a single network structure to map both NIR and VIS images to a compact Euclidean feature space such that the NIR and VIS images in the embedding space directly correspond to the face similarity. The WCNN is composed of three key components in an end-to-end fashion. First of all, inspired by the observations and results indicating that the facial appearance is composed of identity information and variation information (e.g., lightings, poses, and expressions) [18] [4] [19], we divide the high-level layer of the WCNN into two orthogonal subspaces that contain modality-invariant identity information and modality-variant spectrum information. Second, we focus on finding a method for evaluating the distance between the NIR distribution and the VIS distribution. The Wasserstein distance is imposed on the identity subspace to measure the difference between the NIR and the VIS feature distributions, which reduces the gap between the two modalities. The learned identity subspace is expected to contain identity-invariant information of the two modalities. We further assume that the features of the same subject in the identity subspace follow a Gaussian distribution so that the Wasserstein distance can be efficiently optimized. Last, because the fully connected layers of the WCNN have a large number of parameters and are prone to over-fitting on small-scale datasets, we impose a correlation prior on the fully connected layers, which is implemented by a non-convex low-rank constraint. The advantage of this prior is particularly significant when the size of the training set is small.

To make our learned invariant features be robust to heterogeneous intra-class variations of individuals, we first train our convolutional network on large-scale VIS data. We exploit the maxout operator [20] as the activation function in both convolutional layers and fully connected layers. Then, we fine-tune the low-level layers of this network to be adaptable to NIR data. We employ an alternating minimization approach to solve our joint formulation problem, which results in a compact deep representation and an efficient computation for heterogeneous data during training and testing, respectively. The effectiveness of our WCNN method is extensively evaluated using the most challenging CASIA NIR-VIS 2.0 Database [4], Oulu-CASIA NIR-VIS Database [21] and BUAA NIR-VIS Database [22]. Our results demonstrate that the proposed WCNN method clearly outperforms the related state-of-the-art NIR-VIS methods, and significantly improves the state-of-the-art rank-1 accuracy and verification rate (VR) at a low false acceptance rate (FAR).

The main contributions of our work are summarized as follows.

- An effective end-to-end network architecture is developed for learning modality invariant features. This architecture naturally combines invariant feature extraction and subspace learning into a unified network. Two orthogonal subspaces are embedded to model identity and spectrum information, thereby resulting in a single network for extracting both NIR and VIS features.
- A novel Wasserstein distance is introduced to measure the distribution discrepancy between the NIR and the VIS modalities. Compared with previous sample-level measures [16] [9], the Wasserstein distance effectively reduces the gap between the two modalities and results in better feature representation.
- A correlation prior is imposed on the fully connected layers of the deep models to alleviate the over-fitting problem on small scale datasets. This prior improves the performance of the proposed WCNN on a small-scale NIR-VIS dataset and significantly improves the verification rate at a low false acceptance rate.
- The Experimental results on the challenging CASIA NIR-VIS 2.0 face database show that the WCNN increases the best verification rate (@FAR=0.1%) from 91.0% to 98.4%. Compared with the state-of-the-art methods [23], the proposed method further reduces the error rate (1-VR) by 82% with only a compact 128-D feature representation.

The remainder of this paper is organized as follows: Section 2 provides a brief review of related works on NIR-VIS HFR. Section 3 presents the details of our WCNN approach for NIR-VIS face recognition. Section 4 describes the experimental results, and Section 5 summarizes the conclusions.

2 Related Work

The problem of heterogeneous identity matching across different sensing modalities has received increasing attention from the biometrics community, and almost all types of biometrics (e.g., face and iris [24]) have encountered this problem. NIR-VIS HFR represents one of the most extensively researched subject in heterogeneous biometrics. We briefly describe the recent works on this related subject and generally categorize these works into four classes [10] [12] [3]: image synthesis, subspace learning, feature representation and deep learning.

Image synthesis methods aim at transforming face images from one modality (or domain) into another via image synthesis, and they attempt to address the discrepancy between sensing modalities at the image preprocessing stage. These methods were first used in face photo-sketch synthesis and recognition [25], with [26] transforming a face image from one modality to another based on facial analogies and [27] synthesizing a face photo from a pseudo-sketch by

using multi-scale Markov random fields. Then, a hidden Markov model was further used to learn the nonlinear relationship between face photos and sketches [28]. In [6], a canonical correlation analysis (CCA) was used to reconstruct a 3D face model from a single 2D face image; in [29], [30] and [31], face images were reconstructed via coupled or joint dictionary learning and then HFR were performed. Recently, a cross-spectral hallucination and low-rank embedding method was proposed in [32] for synthesizing a VIS image from a NIR image via image patches. Although better rank-1 accuracy was claimed in [32], this work did not follow the standard 10-fold testing protocol [4]. Because image synthesis is an ill-posed problem and a photo-realistic synthesized image is usually difficult to generate, approaches of this type can only reduce the modality difference to a limited extent [3].

Feature representation methods explore modality-invariant features that are robust to various sensing conditions. These methods are often based on hand-crafted local features. Local binary patterns (LBPs), histograms of oriented gradients (HOGs), difference-of-Gaussian (DoG) and scale-invariant feature transform (SIFT) are commonly used [33] [34] [35] [36]. In addition, [37] applied a sparse representation to learn modality-invariant features, and [38] further applied densely sampled SIFT and multi-block LBP features to represent heterogeneous face images. In [10], Log-DoG filtering, local encoding and uniform feature normalization were combined to find better feature representations. Based on the bag of visual words method, [39] proposed hierarchical hyperlingual-words for capturing high-level semantics across different modalities. In [3], face images were converted pixel by pixel into encoded face images with a trained common encoding model, and a discriminant method was applied to match heterogeneous face images. Feature extraction methods reduce the modality difference when converting heterogeneous images to features, and are often applied along with subspace learning methods.

Subspace learning methods aim at learning mappings that project heterogeneous data into a common space in which the inter-modality difference is minimized. Two representative methods are the CCA and partial least squares (PLS) methods. Inspired by the CCA and PLS, [40] proposed a common discriminant feature extraction approach by considering both discriminative and local information. In [41] the locality information was studied in kernel space and a coupled discriminant analysis method was proposed. In [42], a regularized discriminative spectral regression method for projecting heterogeneous data into a common spectral space was developed. Recently, [43] simultaneously performed discriminative feature selection and common subspace learning for different modalities and [44] proposed a prototype random subspace method with kernel similarities for HFR. State-of-the-art NIR-VIS results are often obtained by removing multiple principal subspace components [19]. A multi-view discriminant analysis [45] and mutual component analysis [46] were further developed for reducing the modality difference.

Deep learning methods mainly utilize CNNs to extract deep feature representations of heterogeneous images. These methods are often pre-trained on a large-scale VIS dataset and fine-tuned on NIR face images to learn a modality invariant representation. In [17], a pre-trained VIS CNN was used along with different metric learning strategies to improve the HFR performance. In [16], two types of NIR-VIS triplet loss to reduce intra-class variations were employed and the number of training sample pairs was augmented. In [9], two networks (named VisNet and NIRNet) were trained with small convolutional filters, and the two networks' output features were coupled by creating a Siamese network with contrastive loss. By utilizing the CNN approach, these methods achieved a verification rate of 91.03% at a FAR of 0.1% and a rank-1 accuracy of 95.74% on the challenging CASIA NIR-VIS 2.0 database [16]. However, compared with that of VIS recognition, the performance of NIR-VIS HFR is still far from satisfactory. For example, rank-1 accuracy on the CASIA NIR-VIS 2.0 face database is significantly lower than that on the Labeled Faces in the Wild (LFW) VIS database [47] (rank-1 accuracy is greater than 99%). The high performance of VIS recognition is related to deep learning techniques and the large number of VIS face images. However, because of the gap and the over-fitting problems, NIR-VIS HFR is still challenging for deep learning methods.

The invariant deep representation method was first proposed in our early work [48]. In addition to providing a more in-depth analysis and more extensive experiments, the major difference between this paper and our previous paper [48] is the introduction of the new Wasserstein distance and correlation constraint. Our experiments suggest that the new Wasserstein distance better measures the feature distribution difference between NIR and VIS face data, thereby leading to further improvements in the recognition performance (especially in terms of the FAR). In addition, the correlation constraint on the fully connected WCNN layers increases the adaptability of the learned features to the small-scale NIR training database, which also improves the performance. Compared with our early work [48], our new WCNN method reduces the error rate by 62% at FAR=0.1%.

3 The Proposed Wasserstein CNN

The development of the CNN enabled substantial progress in VIS face recognition in recent years [13] [14] [15]. This section introduces the WCNN, a new CNN architecture for learning modality-invariant deep features for NIR-VIS HFR, and it consists of three key components, as shown in Fig. 1. The first component aims at finding a low-dimensional subspace that contains modality-invariant features. The second component explores the use of the Wasserstein distance to measure the difference between the NIR and VIS distributions. The last component imposes a correlation prior on the fully connected layers to alleviate over-fitting on small-scale NIR datasets.

3.1 Problem Formulation

VIS and NIR images are denoted as I_V and I_N, respectively. The CNN feature extraction process can be formulated as $X_i = Conv(I_i, \Theta_i)$ $(i \in \{N, V\})$, where the feature extraction function $Conv()$ is defined by the ConvNet, X_i indicates a extracted feature vector, and Θ_i is the learned ConvNet parameters for modality I. In HFR, some common concepts are frequently assumed between heterogeneous

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2842770, IEEE Transactions on Pattern Analysis and Machine Intelligence

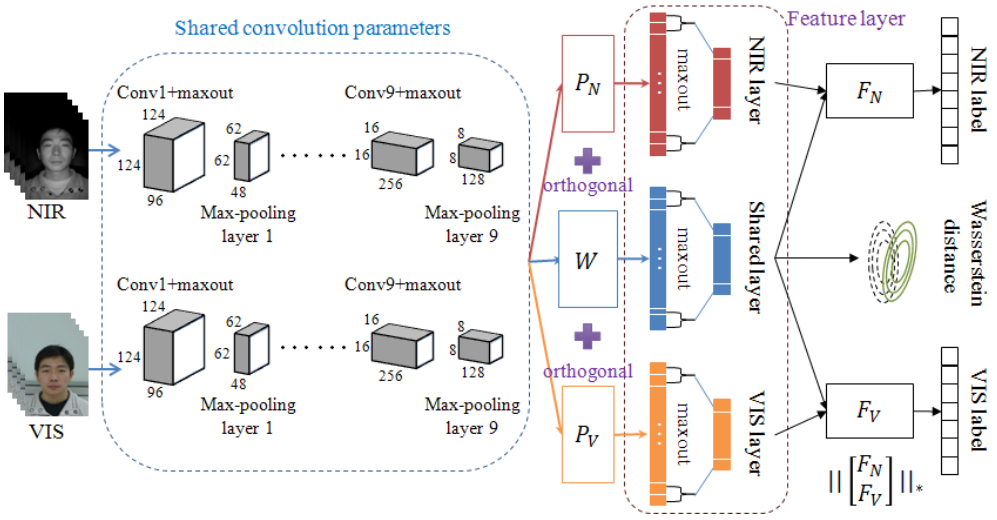


Fig. 1. An illustration of our proposed Wasserstein CNN architecture. The Wasserstein distance is used to measure the difference between NIR and VIS distributions in the modality invariant subspace (spanned by matrix W). At the testing time, both NIR and VIS features are exacted from the shared layer of one single neural network and compared in cosine distance.

modalities. Hence, we also assume that NIR and VIS face images share multiple common low-level features that correspond to the common high-level concepts. Hence, we let $\Theta_N = \Theta_V = \Theta$ and $X_i = Conv(I_i, \Theta)$. As shown in Fig. 1, the output $X_i \in \mathrm{R}^p$ of the last max-pooling layer of the WCNN represents the NIR and VIS channels. These two channels are assumed to share the same parameter Θ.

Modality Invariant Subspace: Previous NIR-VIS matching methods often use a trick to alleviate the problem of appearance variation by removing some principal subspaces that are assumed to contain light spectrum information [4] [19]. Observations and results also indicate that the appearance of a face is composed of identity information and variation information (e.g., lightings, poses, and expressions) [18] and removing spectrum information is helpful for the NIR-VIS performance [19]. Inspired by these results, we introduce three mapping matrices (i.e., $W, P_i \in \mathrm{R}^{d\times p}$) in the CNN for modeling identity invariant information and variant spectrum information. Therefore, the deep feature representation can be defined as follows:

$$f_i = \begin{bmatrix} f_{\text{shared}} \\ f_{\text{unique}} \end{bmatrix} = \begin{bmatrix} WX_i \\ P_iX_i \end{bmatrix} \ (i \in \{N, V\}), \tag{1}$$

where WX_i and P_iX_i denote the shared feature and the unique feature respectively. Considering the subspace decomposition properties of the matrices W and P_i, we further impose an orthogonality constraint to make them unrelated to each other:

$$P_i^T W = 0 \ (i \in \{N, V\}). \tag{2}$$

This orthogonality constraint can also reduce the parameter space and alleviate over-fitting. Compared with previous methods [19] [46] [39] [3] that treat feature representation and subspace learning as two independent steps, our architecture naturally combines these two steps into an end-to-end network.

The Wasserstein Distance: The gap in the sensing mechanism between NIR and VIS images is a major difficulty in HFR. Previous methods often resort to sample-level constraints to reduce this gap. The triplet loss and contrastive loss constraints are imposed on NIR-VIS sample pairs in [11] and [9] respectively. These methods only consider the relationships among NIR-VIS samples rather than between NIR-VIS distributions. Recently, the Wasserstein distance was shown to play a prominent role in measuring the model distribution and the real distribution in generative adversarial networks (GAN) [49] [50]. Inspired by the Wasserstein GAN [49] and BEGAN [50], we use the Wasserstein distance to measure the consistency between the NIR and the VIS data distribution. Because NIR-VIS data are from different subjects and large extra-class variations are observed, we impose the Wasserstein distance on the distributions of a single subject. We further assume that the data distributions of the same subject follow a Gaussian distribution after nonlinear feature mapping. The Gaussian distribution assumption of the Wasserstein distance has been shown to be effective in the image generation problem [50] and the sequence matching problem [51]. The experimental results show that this assumption also provides meaningful learning results for HFR.

Given two Gaussian distributions $X = \mathcal{N}(m_N, C_N)$ and $Y = \mathcal{N}(m_N, C_N)$ that correspond to one subject, where the means are $m_N, m_V \in \mathbb{R}^p$ and the covariances are $C_N, C_V \in \mathbb{R}^{p\times p}$, the 2-Wasserstein distance between X and Y of one subject is defined as follows [50]:

$$W_2(X,Y)^2 = \|m_N - m_V\|_2^2 + \text{trace}(C_N + C_V - 2(C_V^{\frac{1}{2}} C_N C_V^{\frac{1}{2}})^{\frac{1}{2}}). \tag{3}$$

As in [50], we simplify (3) to:

$$W_2(X,Y)^2 = \frac{1}{2}\left[\|m_N - m_V\|_2^2 + \|\sigma_N - \sigma_V\|_2^2\right], \tag{4}$$

where σ_N and σ_V are the standard deviations of X and Y, taking the following forms:

$$\sigma_1 = \sqrt{\frac{1}{n}\sum_{i=0}^{n} x_i^2 - m_N^2}, \quad \sigma_2 = \sqrt{\frac{1}{n}\sum_{i=0}^{n} y_i^2 - m_V^2}. \tag{5}$$

The gradient of X can be computed as

$$\frac{\partial W_2}{\partial x_i} = \frac{1}{n}\left[(m_N - m_V) + 2(\sigma_N - \sigma_V)\frac{(x_i - m_N)}{\sqrt{\sigma_N^2 + \epsilon}}\right], \quad (6)$$

where ϵ is a constant. Analogously, the gradient of Y can be written as

$$\frac{\partial W_2}{\partial y_i} = -\frac{1}{n}\left[(m_N - m_V) + 2(\sigma_N - \sigma_V)\frac{(y_i - m_V)}{\sqrt{\sigma_V^2 + \epsilon}}\right]. \quad (7)$$

Correlation Prior: One challenge of applying a CNN to HFR is the over-fitting problem of CNNs on a small-scale training set. In a CNN, fully connected layers often contribute the majority of the parameters. Because both NIR and VIS labels are used in HFR, the number of class labels in HFR is twice as large as that in VIS face recognition. A large number of class labels also result in fully connected layers of a large size. Hence, when the training set is of a small-scale, fully connected layers can not be well adjusted and may easily over-fit the data. The fully connected layer of the WCNN is composed of two matrices, which are denoted as F_N and F_V and correspond to the NIR and the VIS modalities, respectively. We expected $\mathrm{M} = \begin{bmatrix} F_N \\ F_V \end{bmatrix}$ to be highly correlated such that $M^T M$ is a block-diagonal matrix [1]. A correlated M will reduce the size of the estimated parameter space and naturally alleviate the over-fitting problem. We exploit the matrix nuclear norm on M, i.e.,

$$\|\mathrm{M}\|_* = \mathrm{tr}(\sqrt{\mathrm{M}^T\mathrm{M}}). \quad (8)$$

The matrix nuclear norm requires that M has a low-rank structure and its elements are linearly correlated. Then $M^T M$ tends to be a block-diagonal matrix. Given the SVD decomposition of $\mathrm{M} = U\Sigma V^T$, we can obtain:

$$\mathcal{R} = \|\mathrm{M}\|_* = \mathrm{tr}(\sqrt{V\Sigma U^T U\Sigma V^T}) = \mathrm{tr}(\sqrt{\Sigma^2}). \quad (9)$$

Since the elements of Σ are non-negative, the gradient of the nuclear norm can be written as:

$$\frac{\partial \mathcal{R}}{\partial \mathrm{M}} = \frac{\partial \mathrm{tr}(\Sigma)}{\partial \mathrm{M}} = UV^T. \quad (10)$$

Therefore, we can use UV^T as the subgradient of the nuclear norm. Note that because the fully connected matrices F_V and F_N are not used for testing, the correlation prior is only intended to alleviate over-fitting and not to compress the network.

3.2 Optimization Method

The commonly used softmax loss is used to train the whole network, taking the following form,

$$\begin{aligned}\mathcal{L}_{cls} &= \sum_{i\in\{N,V\}} \mathrm{softmax}(F_i, c, \Theta, W, P_i) \\ &= -\sum_{i\in\{N,V\}} (\sum_{j=1}^{N} \mathbf{1}\{y_{ij} = c\} \log \hat{p}_{ij}) \\ \text{s.t.} \quad & P_i^T W = 0 \ (i \in \{N, V\})\end{aligned} \quad (11)$$

1. Block-diagonal prior was used in subspace segmentation to make clustering results more accurately [52]. It requires an affinity matrix to be block-diagonal to characterize sample clusters

Algorithm 1: Training the Wasserstein CNN network.

Require: Training set X_i, learning rate γ and lagrange multipliers λ_i.
Ensure: The CNN parameters Θ and the mapping matrix W.
1: Initialize parameters Θ by pre-trained model and the mapping matrices W, P_i, F_i by Eq.(24);
2: **for** $t = 1, \ldots, T$ **do**
3: CNN optimization:
4: Update Θ, W, P_i, F_i via back-propagation method;
5: Fix Θ:
6: Update W according to Eq.(18);
7: Update P_i according to Eq.(19);
8: Update F_i according to Eq.(10);
9: **end for**;
10: **Return** Θ and W;

where c is the class label for each sample and $\hat{p}_{ij}$ is the predicted probability. Moreover, we denote $\mathbf{1}\{\cdot\}$ as the indicator function so that $\mathbf{1}\{\text{a true statement}\} = 1$ and $\mathbf{1}\{\text{a false statement}\} = 0$.

According to the theory of lagrange multipliers, (14) can be reformulated as an unconstrained problem,

$$\begin{aligned}\mathcal{L}_{cls} = &\sum_{i\in\{N,V\}} \mathrm{softmax}(F_i, c, \Theta, W, P_i) \\ &+ \sum_{i\in\{N,V\}} \lambda_i \|P_i^T W\|_F^2,\end{aligned} \quad (12)$$

where λ_i are the lagrange multipliers and $\|\cdot\|_F^2$ denotes the Frobenius norm.

To decrease the discrepancy between different modalities, we apply the Wasserstein distance to measure two distributions of NIR and VIS images from the same subject.

$$\mathcal{L}_{dist} = \frac{1}{2}\left[\|m_N - m_V\|_2^2 + \|\sigma_N - \sigma_V\|_2^2\right]. \quad (13)$$

Under the WCNN training scheme, we employ mini-batch stochastic gradient descent to optimize the objective function, and use the statistics of each mini-batch to represent the means and standard deviations instead.

To alleviate over-fitting, we also introduce Eq.(9). Then the final objective function takes the following form,

$$\mathcal{L} = \beta_1 \mathcal{L}_{cls} + \beta_2 \mathcal{L}_{dist} + \beta_3 R, \quad (14)$$

where β_1, β_2 and β_3 are the trade-off coefficients for each part. If the gradient descent method is used to minimize Eq.(14), we should update parameters W, P_i, F_i and Θ. We follow the back-propagation method to update the vector of convolutional parameters Θ. The gradients of W, P_i and F_i can be expressed as

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}_{cls}}{\partial W} + \frac{\partial \mathcal{L}_{dist}}{\partial W} \quad (15)$$

$$\frac{\partial \mathcal{L}}{\partial P_i} = \frac{\partial \mathcal{L}_{cls}}{\partial P_i} + \frac{\partial \mathcal{L}_{dist}}{\partial P_i} \quad (16)$$

$$\frac{\partial \mathcal{L}}{\partial F_i} = \frac{\partial \mathcal{L}_{cls}}{\partial F_i} + \frac{\partial \mathcal{R}}{\partial F_i} \quad (17)$$

Note that the updating gradients for W, P_i and F_i contain two parts. The first one is used for conventional back-propagation in CNN. The second part of W, P_i for subspace learning can be re-organized in

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{i \in \{N,V\}} \lambda_i P_i P_i^T W \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial P_i} = \lambda_i W W^T P_i \tag{19}$$

For the low-rank correlation constraint, we can update $M = [F_N, F_V]^T$ by Eq.(10). Then we update these parameters with a learning rate γ via

$$\Theta^{(t+1)} = \Theta^{(t)} - \gamma \frac{\partial \mathcal{L}}{\partial \Theta^{(t)}} \tag{20}$$

$$W^{(t+1)} = W^{(t)} - \gamma \frac{\partial \mathcal{L}}{\partial W^{(t)}} \tag{21}$$

$$P_i^{(t+1)} = P_i^{(t)} - \gamma \frac{\partial \mathcal{L}}{\partial P_i^{(t)}} \tag{22}$$

$$F_i^{(t+1)} = F_i^{(t)} - \gamma \frac{\partial \mathcal{L}}{\partial F_i^{(t)}} \tag{23}$$

Because Eq.(14) contains several variables and is non-convex, we develop an alternating minimization method for minimizing Eq.(14) in an end-to-end CNN optimization scheme. First, we update the parameters by conventional back-propagation to optimize the CNN. Then, we fix the CNN parameters and update matrices W, P_i, F_i by their own gradients. The optimization details are summarized in Algorithm 1. As in [53], the vector of parameters Θ of the CNN is initialized by the pre-trained model and the mapping matrices W, P_i, F_i are initialized by

$$W, P_i, F_i \sim U\left[-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\right] \tag{24}$$

where $U[-a, a]$ is the uniform distribution in the interval $(-a, a)$ and m is the dimension of the original features.

3.3 Network Structure

The basic VIS network (the convolutional parameter sharing part in Fig. 1) is built based on the light CNN network[2] [54]. The network is composed of nine convolutional layers with four max-pooling layers, and then a fully connected layer. The values of Θ are initialized by training the light CNN on a large-scale VIS dataset [55], and the Softmax and maxout operations are used as the loss function and activation function, respectively. The MS-Celeb-1M dataset [55], which contains a total of 8.5M images of approximately 100K subjects, is used for training. We normalize and crop all training VIS images to 144×144 according to five facial points. To enrich the training data, we randomly crop the training images to 128×128. We set the dropout ratio to 0.7 for the fully connected layer. We initialize the learning rate to $1e^{-3}$ and gradually reduce it to $1e^{-5}$ over $4,000,000$ iterations. The trained single VIS network obtained an accuracy of 98.90% on the challenging LFW dataset.

2. https://github.com/AlfredXiangWu/face_verification_experiment

Using the pre-trained basic VIS network, we further learn a modality-invariant Wasserstein CNN from the NIR-VIS data. We initialize the low-level convolution layers of the WCNN via the pre-trained basic VIS network. NIR and VIS images are input into two CNN channels with shared parameters. Then, a feature layer (as in shown Fig. 1) is defined for projecting the low-level features into two orthogonal feature subspaces. Therefore, the correlated properties of the NIR and VIS identities can be leveraged and the domain-specific properties of both modalities are enforced. When the sum of the Wasserstein distances over all subjects reaches zero, invariant deep features are learned. Finally, the Softmax loss functions are separately used for the NIR and VIS representations as the supervisory signals. Because a maxout operator is included in the feature layer, the final feature dimension is $d/2$ when $W \in \mathrm{R}^{d \times m}$. As in the VIS training, we crop and resize all NIR and VIS images to 144×144 pixels and randomly select a region of size 128×128 pixels for NIR-VIS training. The learning rate of the WCNN is set to $1e^{-4}$ initially and reduced to $1e^{-6}$ gradually for approximately $100,000$ iterations. The trade-off parameters β_1, β_2, β_3 and λ_i are set to 1, 1, 0.001, and 0.001, respectively. Because the numerical value of each term in Eq. (14) is significantly different, their trade-off parameters are simply set such that the value of each term is on the same scale.

4 Experiments and Results

In this section, we systemically evaluate the proposed WCNN approach against traditional methods and deep learning methods on three recently published NIR-VIS face databases: the CASIA NIR-VIS 2.0 database, the Oulu-CASIA NIR-VIS database and the BUAA-VisNir database. Fig. 2 shows samples of cropped VIS and NIR facial images from the three databases.

4.1 Datasets and Protocols

CASIA NIR-VIS 2.0 Face Database: this database [4] is the largest public and most challenging NIR-VIS heterogeneous face database. This database contains images of the same subjects with large variations, including lighting, expression, pose, and distance. To simulate real-world variations, the database also contains images of subjects wearing glasses or not. This database was collected in four sessions from 2007 to 2010. A total of 725 subjects is included, and they range from children to the elderly. Each subject has 1-22 VIS and 5-50 NIR images. Each image is randomly gathered without a one-to-one correlation between NIR and VIS images. Two types (views) of evaluation protocols are used for super-parameter adjustments and training/testing.

We follow the standard protocol in View 2 to evaluate the different methods. Ten-fold experiments are performed, with each considering a collection of training and testing lists. The training and testing sets contain nearly equal numbers of subjects. For each training fold, NIR and VIS training images are randomly combined. To form approximately 2,500 VIS images and 6,100 NIR images from approximately 360 subjects. These subjects are different from the 358 subjects in the testing set. Thus, the subjects in the training

(a) The CASIA NIR-VIS 2.0 database (b) The Oulu-CASIA NIR-VIS database (c) The BUAA-VisNir database

Fig. 2. Cropped VIS and NIR facial images from the three databases. The first row contains the NIR images from the probe set and the second row contains the VIS images from the gallery set.

set and the testing set are different and the two sets are disjointed. We train the WCNN on the training set in each fold. For each testing fold, the gallery set always contains a total of 358 VIS images from 358 subjects. That is, each subject only has one VIS image. The probe set is composed of over 6,000 NIR images from the same 358 subjects. Each NIR image in the probe set is matched against each VIS image in the gallery set, which results in a similarity matrix of size 358 by approximately $6,000$.

Oulu-CASIA NIR-VIS database: this database [21] is composed of 80 subjects with six expression variations (anger, disgust, fear, happiness, sadness, and surprise). Fifty subjects are from Oulu University, and the remaining 30 subjects are from CASIA. Because the facial images of this database are captured under different environments from two institutes, their illumination conditions are slightly different [39]. Following the protocols in [39], we select a subset of this database for our experiments, which includes 10 subjects from Oulu University and 30 subjects from CASIA. Eight face images for each expression are randomly selected from each of the NIR and VIS datasets. As a result, a total of 96 images (48 NIR images and 48 VIS images) are available for each subject. Twenty subjects are used for training, and the remaining 20 subjects are used for testing. All VIS images of the 20 subjects that are used for testing constitute the gallery set, and all their corresponding NIR images constitute the probe set.

BUAA-VisNir face database: this database [22] is often used for domain adaptation evaluations across imaging sensors. This database has contains images of 150 subjects, with 9 VIS images and 9 NIR images of each subject captured simultaneously. The nine images of each subject correspond to nine distinct poses or expressions: neutral-frontal, left-rotation, right-rotation, tilt-up, tilt-down, happiness, anger, sorrow and surprise. The training set and testing set are composed of 900 images of 50 subjects and 1800 images of the remaining 100 subjects respectively. As in [21], to avoid the probe and gallery images showing in the same pose and expression, only one VIS image of each subject is selected from the gallery set during testing. Hence, the gallery set and probe set have 100 VIS images and 900 NIR images, respectively. This testing protocol is challenging due to the large pose and illumination variations in the probe set.

4.2 Results for the CASIA NIR-VIS 2.0 Database

We demonstrate the effectiveness of the proposed WCNN by comparing it with state-of-the-art methods, including traditional methods and deep learning methods. Because most methods follow the standard protocol, to evaluate their performance for the CASIA NIR-VIS 2.0 database, we directly report their results from the published papers.

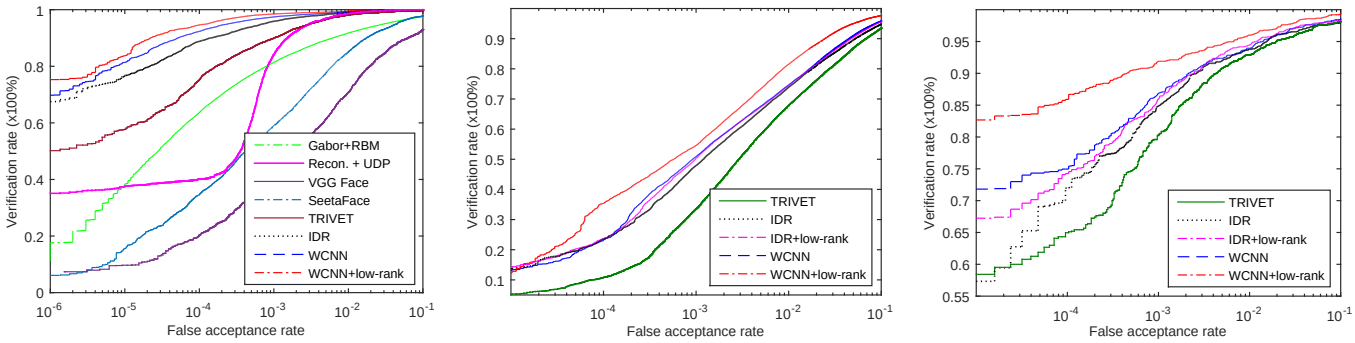
TABLE 1
Experimental results on the 10-fold CASIA NIR-VIS 2.0 benchmark in terms of rank-1 accuracy, verification rate and standard deviation.

Methods	Rank-1	FAR=1%	FAR=0.1%	Dim
KCSR [56]	33.8	28.5	7.6	-
KPS [44]	28.2	17.4	3.7	-
KDSR [42]	37.5	33.0	9.3	-
PCA+HCA [4]	23.7±1.9	-	19.3	-
LCFS [43] [12]	35.4±2.8	35.7	16.7	-
H2(LBP3) [39]	43.8	36.5	10.1	-
C-DFD [57] [12]	65.8±1.6	61.9	46.2	-
DSIFT [58]	73.3±1.1	-	-	-
CDFL [12]	71.5±1.4	67.7	55.1	1000
Gabor+RBM [19]	86.2±1.0	-	81.3±1.8	-
Recon.+UDP [31]	78.5±1.7	-	85.8	1024
CEFD [3]	85.6	-	-	-
VGG [59]	62.1±1.9	70.9±1.3	39.7±2.9	4096
SeetaFace [23]	68.0±1.7	85.2±1.1	58.8±2.3	2048
TRIVET [16]	95.7±0.5	98.1±0.3	91.0±1.3	512
HFR-CNNs [17]	85.9±0.9	-	78.0	-
IDNet [9]	87.1±0.9	-	74.5	320
IDR [48]	97.3±0.4	98.9±0.3	95.7±0.7	128
WCNN	98.4±0.4	99.4±0.1	97.6±0.3	128
WCNN+low-rank	**98.7**±0.3	**99.5**±0.1	**98.4**±0.4	128

The traditional methods include kernel coupled spectral regression (KCSR) [56], kernel prototype similarities (KPS) [44], kernel discriminative spectral regression (KDSR) [42], PCA+HCA [4], learning coupled feature spaces (LCFSs) [43], coupled discriminant face descriptor (C-DFD) [57], D-SIFT+PCA+LDA [58], coupled discriminant feature learning (CDFL) [12], Gabor+RBM+Remove 11PCs [19], reconstruction+UDP [31], H2(LBP3) [39], common encoding feature discriminant (CEFD) [3]. The results of the LCFS, C-DFD and CDFL methods are from [12], and those of the remaining methods are from their respective published papers. For the deep learning methods, we compare the recently proposed TRIVET [16], HFR-CNNs [17] and IDNet [9]. In addition, the results of the following two VIS CNN methods are discussed: VGG [59] and SeetaFace [23].

Table 1 shows the rank-1 accuracy and verification rates of different NIR-VIS methods. Fig. 3 (a) further plots the receiver operating characteristic (ROC) curves of the proposed method and its three top competitors. To improve

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2842770, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) The CASIA NIR-VIS 2.0 database (b) The Oulu-CASIA NIR-VIS database (c) The BUAA-VisNir database

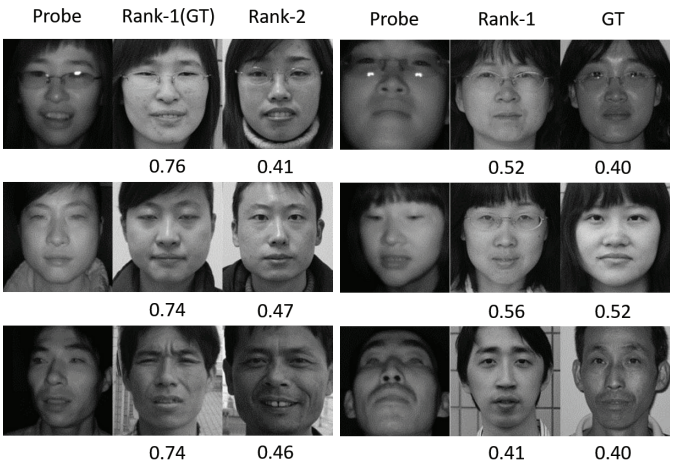
Fig. 3. ROC curves of different methods on the three NIR-VIS datasets.

the illustration, we do not report the ROC curves of other methods if these curves are low. The following observations were recorded:

Because of the sensing gap, the three VIS deep models do not perform well for NIR-VIS HFR. The rank-1 accuracy and VR@FAR=0.1% of the VGG and SeetaFace methods are lower than those of the state-of-the-art traditional methods, and significantly worse than those of the deep learning methods trained on the NIR-VIS dataset. Compared with VGG and SeetaFace, the CEFD and Gabor+RBM methods can also obtain higher rank-1 accuracy. These results suggest that although the large-scale VIS dataset is helpful for VIS face recognition, it has limited benefits for HFR if only a small-scale NIR dataset is available. Hence a suitable deep structure for NIR and VIS modalities must be designed. Then the deep learning based methods (TRIVET, HFR-CNNs and IDNet) may outperform the traditional methods.

Compared with the capabilities of traditional methods (CEFD, Gabor+RBM and reconstruction+UDP), the improvements achieved by the recently proposed deep learning based methods (TRIVET, HFR-CNNs and IDNet) are limited. In particular, high rank-1 accuracy does not ensure a high verification rate or a better ROC curve. The experimental results clearly show that our WCNN method yields a superior overall performance compared with that of the other NIR-VIS methods. One of the main strengths of the WCNN is its ability to yield consistent improvements in terms of the rank-1 accuracy and verification rates. The advantage of the WCNN is particularly apparent when the FAR is low. Moreover, the standard deviations of the WCNNs are also smaller than those of their competitors. The smaller standard deviations indicate that the WCNNs obtain better results on all 10-fold experiments. In particular, because we utilize an orthogonal subspace to separate spectral information and identity information, the feature dimension of our method is smaller than that of the other methods. All these results suggest that deep learning is effective for the NIR-VIS recognition problem, and a compact and modality-invariant feature representation can be learned from a single CNN.

Compared with our earlier version (IDR) [48], the WCNN+low-rank method further improves the rank-1 accuracy from 97.3% to 98.7% and VR@FAR=0.1% from 95.7% to 98.4%. It further reduces the error rate (1-VR) by 62% at FAR=0.1%. Although the rank-1 accuracy and VR@FAR=0.1% of the WCNN are high, the low-rank constraint could still improve the performance of the WCNN. Note that 2,148,000 NIR-VIS pairs are included in the testing set. Hence, a small improvement will result in the correct classification of many NIR-VIS pairs. These results highlight the importance of the Wasserstein distance and the low-rank constraint for the sensing gap and over-fitting problems. When these two problems are well treated, deep learning methods significantly improve the NIR-VIS recognition performance.



(a) Correctly matching results (b) Failed cases

Fig. 4. Visual results of correctly matched faces and failed cases on the CASIA NIR-VIS 2.0 database. The numerical score under each gallery image is its similarity score with the probe image in the same row.

Fig. 4 shows the visual results of correctly matched faces and failed cases in terms of the rank-1 accuracy. The left image is from the probe set, and the right two images are from the gallery set. The numerical score under each gallery image is its similarity score with the probe image in the same row. The notation 'GT' indicates that a gallery image and a probe image correspond to the same subject. From the correctly matched results in Fig. 4 (a), we observe that a large margin occurs between the ground-truth scores ('GT'), and the impostor scores ('Rank-2'). The ground-truth scores are all larger than 0.7, and the impostor scores are all

TABLE 2
Rank-1 accuracy and verification rate on the Oulu-CASIA NIR-VIS Database.

Methods	Rank-1	FAR=1%	FAR=0.1%
MPL3 [21]	48.9	41.9	11.4
KCSR [56]	66.0	49.7	26.1
KPS [44]	62.2	48.3	22.2
KDSR [42]	66.9	56.1	31.9
H2(LBP3) [39]	70.8	62.0	33.6
TRIVET [16]	92.2	67.9	33.6
IDR	94.3	73.4	46.2
IDR+low-rank	95.0	73.6	50.3
WCNN	96.4	75.0	50.9
WCNN + low-rank	**98.0**	**81.5**	**54.6**

TABLE 3
Rank-1 accuracy and verification rate on the BUAA NIR-VIS Database.

Methods	Rank-1	FAR=1%	FAR=0.1%
MPL3 [21]	53.2	58.1	33.3
KCSR [56]	81.4	83.8	66.7
KPS [44]	66.6	60.2	41.7
KDSR [42]	83.0	86.8	69.5
H2(LBP3) [39]	88.8	88.8	73.4
TRIVET [16]	93.9	93.0	80.9
IDR	94.3	93.4	84.7
IDR + low-rank	94.8	94.5	86.0
WCNN	95.4	93.9	86.9
WCNN + low-rank	**97.4**	**96.0**	**91.9**

smaller than 0.5. This large margin also enables our WCNNs to obtain better ROC curves. Fig. 4 (b) shows the failed cases in which the rank-1 gallery image with the highest score does not correspond to the same subject as the probe image. We observe that the ground-truth scores ('GT') and the impostor scores ('Rank-1') are similar, and they are all smaller than 0.6. These failed cases may be related to large pose variations. Both the pose variations and the sensing gap lead to large appearance variations, which increases the difficulty of identifying the probe images in Fig. 4 (b), even by the human eye.

4.3 Results for the Oulu-CASIA NIR-VIS Database

In this subsection, we evaluate the proposed methods for the Oulu-CASIA NIR-VIS database. Compared with that of the CASIA NIR-VIS 2.0 database, the training set of the Oulu-CASIA NIR-VIS database only consists of 20 subjects, which represents a relatively small-scale. Hence, this database is challenging for a deep learning method because of over-fitting. We follow the testing protocol in [39] and compare the WCNN with the MPL3 [21], KCSR [56], KPS [44], KDSR [42], KDSR [42], H2(LBP3) [39] and TRIVET [16]. The results of the MPL3, KCSR, KPS, KDSR, KDSR and H2(LBP3) methods are from [39]. The TRIVET method is used as the baseline of deep learning methods.

Table 2 shows the rank-1 accuracy and verification rates of different NIR-VIS matching methods. We observe that the methods can be ordered according to the ascending rank-1 accuracy as follows: MPL3, KPS, KCSR, KDSR, H2(LBP3), TRIVET, IDR, WCNN and WCNN+low-rank. The four deep learning methods perform significantly better than the five traditional methods in terms of rank-1 accuracy. Although the rank-1 accuracy of TRIVET is higher than that of H2(LBP3), the VR@FAR=0.1% value of TRIVET is close to that of H2(LBP3). This finding may be because all VIS images of each subject are from the gallery and all their corresponding NIR images are treated as probes. Because a NIR image is paired with a VIS image during testing, a deep learning method can easily obtain a high similarity score for paired data. Thus, the rank-1 accuracy of one deep learning method is high. However, due to the sensing gap, the NIR image feature of one person may be similar to the VIS image feature of another person showing the same expression. These two features may also have a higher similarity score. Therefore, the verification rates of all methods are not high at a low FAR. Because of the small-scale of the training set for this database, the four deep learning methods can not capture all variations. Thus, their verification rates are lower than those for the CASIA NIR-VIS 2.0 database. As expected, the WCNN method achieves the highest performance in terms of rank-1 accuracy and verification rate.

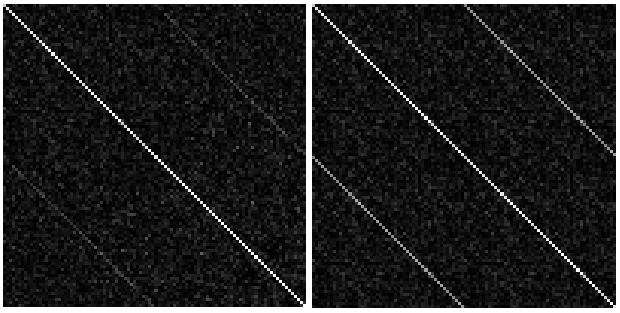
Fig. 3 (b) plots the ROC curves of the four deep learning methods. The verification rates of all four methods drop dramatically as FAR becomes small. TRIVET obtains the lowest ROC curve. Only small improvement is observed between the curves of the WCNN and IDR. When the low-rank constraint is imposed on IDR, the ROC curve of IDR+low-rank is close to that of the WCNN, which means that Wasserstein distance does not contribute substantially to the ROC curve. This finding is mainly because the training set of this database is of a small-scale. As a result, the WCNN over-fits this small-scale training set. When the low-rank constraint is imposed on the fully connected layer of the WCNN, a significant difference is observed between the ROC curves of the WCNN and WCNN+low-rank. These results suggest that a suitable constraint on the fully connected layer can alleviate the over-fitting problem on small training sets.

4.4 Results for the BUAA VisNir Database

In this subsection, we evaluate the proposed methods on the BUAA VisNir database. As shown in Fig. 2 (c), the VIS and NIR images are well aligned and have a similar appearance because they are captured simultaneously. These well-aligned NIR and VIS images potentially facilitate the capture of intrinsic identity variations for deep learning methods and reduce the sensing gap. We follow the testing protocol in [39] to evaluate the NIR-VIS matching methods. The results for the BUAA VisNir database are presented in Table 3 and Fig. 3 (c). The results of the MPL3, KCSR, KPS, KDSR, KDSR and H2(LBP3) methods are from [39].

We observe that the five deep learning methods perform better than the five traditional methods. The methods can be ordered according to the ascending rank-1 accuracy as MPL3, KPS, KCSR, KDSR, H2(LBP3), TRIVET, IDR, IDR+low-rank, WCNN and WCNN+low-rank. Our WCNN+low-rank method improves the best rank-1 accuracy from 88.8% to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2842770, IEEE Transactions on Pattern Analysis and Machine Intelligence



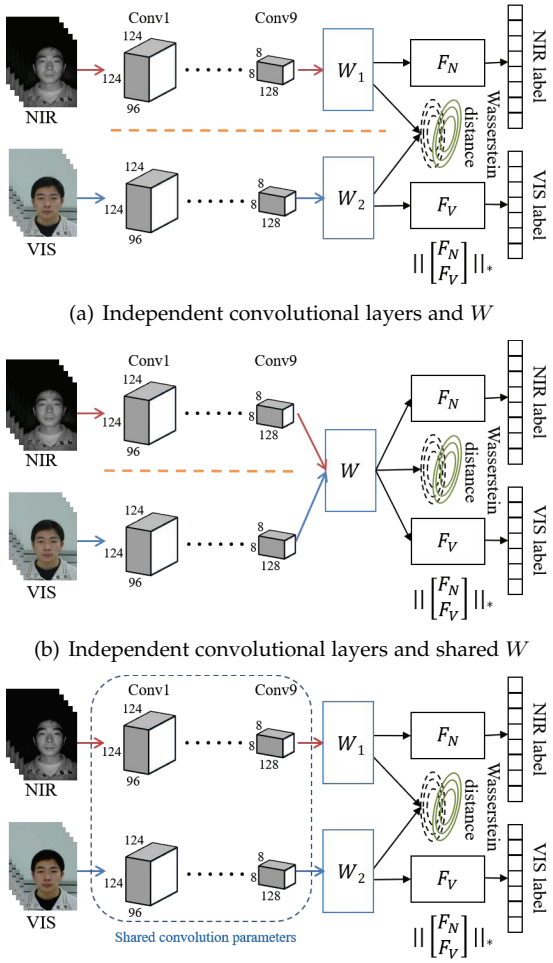
(a) Without low-rank constraint (b) With low-rank constraint

Fig. 5. A correlation illustration of the matrix $\mathrm{M}^T\mathrm{M}$ in the fully connected layer of WCNN. A lighter color indicates a higher correlation. When the low-rank correlation constraint is introduced, there is obvious variations on top-right and bottom-left areas of $\mathrm{M}^T\mathrm{M}$.

97.4% and VR@FAR=0.1 from 73.4% to 91.9%. When the low-rank constraint and the Wasserstein distance are introduced to the IDR, the IDR's performance is significantly improved. In particular, the highest performance is achieved when both the low-rank constraint and the Wasserstein distance are used because deep learning methods are simultaneously degraded by the sensing gap and over-fitting problems. Our proposed architecture can naturally address these two problems in an end-to-end network, thereby resulting in higher performance for this database.

We also observe that WCNN+low-rank can further improve the performance of IDR+low-rank, suggesting that Wasserstein distance does work on this database. Note that our WCNNs aim to reduce the discrepancy between NIR and VIS distributions. Although there are only nine images per subject on the BUAA VisNir dataset and distribution estimation may be inaccurate, WCNNs can still find a feature space in which the discrepancy between NIR and VIS distributions is minimized. Generally, the more images per subject are used, the better the discrepancy between two distributions can be obtained. Due to limited training samples, the performance of WCNNs on the BUAA VisNir database is lower than that on the CASIA NIR-VIS 2.0 database.

Fig. 3 (c) shows that the methods can be ordered according to the ascending ROC curves as follows, TRIVET, IDR, IDR+low-rank, WCNN and WCNN+low-rank. The low-rank constraint significantly improves the ROC curves of the IDR and WCNN, especially when the FAR is low. Because the training set of this database is of a small-scale, deep learning methods may over-fit the training set. Fig. 5 plots the values of matrix $\mathrm{M}^T\mathrm{M}$ without (Fig. 5 (a)) and with (Fig. 5 (b)) the low-rank constraint on the fully connected layer of the WCNN. A lighter color indicates a higher correlation. When the low-rank correlation constraint is used, obvious variations occur in the top-right and bottom-left areas of the $\mathrm{M}^T\mathrm{M}$. Note that M is composed of F_N and F_V. The diagonal elements in the top-right and bottom-left areas have a lighter color, which indicates that F_N and F_V are correlated, thus reducing the size of the parameter space of the fully connected layer. These results further validate the effectiveness of the low-rank correlation constraint, which suggests that correlation constraints can be used on the fully connected layer to alleviate the over-fitting problem.



(a) Independent convolutional layers and W

(b) Independent convolutional layers and shared W

(c) Shared convolutional layers and Independent W

Fig. 6. Three alternative architectures of WCNN.

4.5 Ablation Study

In this subsection, we investigate different architectures and configurations of our proposed WCNN method via the 10-fold evaluation on the CASIA NIR-VIS 2.0 database. Six variations of the WCNN are considered: three for comparing architectures and three for comparing objective functions. As in our previous conference version [48], the notation 'DR' indicates the basic network of the WCNN with only the shared convolutional layers and the shared mapping matrix W, in which the parameters β_3, β_2 and λ_i are all set to zero. IDR indicates the variation of the WCNN without the simplified Wasserstein distance. We also study the effect of different convolutional layers and mapping matrices (or transformations) for the WCNN. In Fig. 6 (a), (b) and (c), we illustrate the three variants of the WCNN as WCNN1+low-rank, WCNN2+low-rank and WCNN3+low-rank, respectively. In WCNN1+low-rank, two different convolutional layers and mapping matrices are used for NIR and VIS images respectively. In the WCNN2+low-rank method, two different convolutional layers and a shared mapping matrix are considered. In the WCNN3+low-rank method, shared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2842770, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 4
Performance comparison of different WCNN variants.

Methods	Setting	Rank-1	FAR=0.1%
DR	$\beta_3 = \beta_2 = \lambda_i = 0$	93.8±0.5	91.8±0.8
IDR	$\beta_3 = \beta_2 = 0$	97.3±0.4	95.7±0.7
WCNN	$\beta_3 = 0$	98.4±0.4	97.6±0.3
WCNN1+low-rank	-	96.9±0.3	97.3±0.4
WCNN2+low-rank	-	97.7±0.3	97.1±0.7
WCNN3+low-rank	-	98.5±0.3	98.0±0.3
WCNN+low-rank	-	98.7±0.3	98.4±0.4

convolutional layers and two different mapping matrices are considered.

Table 4 tabulates the comparison results of the six WCNN variants. Each component of Eq. (14) makes an important contribution to the WCNN in terms of both rank-1 accuracy and verification rate. When applying CNNs to NIR-VIS HFR, one should address the sensing gap and over-fitting problems. Each component of the WCNN is elaborately integrated into an end-to-end CNN network to address these two challenging problems, which results in significant improvements in the recognition performance. The effectiveness of the Wasserstein distance and the low-rank constraints of the WCNN has also been verified in Table 3 and Table 2.

The results show that the WCNN1+low-rank and WCNN2+low-rank methods do not yield better results than the WCNN+low-rank method because they have larger convolutional parameter spaces, and large-scale NIR-VIS training images are not available. The shared strategy in the WCNN actually provides a method of constraining the convolutional parameter space of the WCNN. Comparing with the WCNN3+low-rank and WCNN+low-rank methods, we observe that there is also a performance drop of WCNN3+low-rank. This finding further demonstrates that the architecture of the orthogonal subspace in the WCNN+low-rank method performs better than the architecture of the two mapping matrices in the WCNN3+low-rank method. In particular, during the deep feature extraction in testing, the WCNN+low-rank method only needs one mapping matrix (P_N and P_V are not used during testing) whereas the WCNN3+low-rank method needs two mapping matrices.

Moreover, in heterogenous face recognition, the major discrepancy between NIR and VIS images is the light (or sensing) spectral information, the NIR and VIS images from the same person have the unique identity information. If deep convolution parameters are powerful enough to remove spectral information from identity information, one can use the same parameters for NIR and VIS images to extract identity information. Particularly, there are three projection matrices after convolution parameters. P_N, P_V and W correspond to NIR spectral information, VIS spectral information and identity information respectively. If spectral information is well separated, the projection matrix W only contains the same identity information. Hence, one can potentially use the identity related subnetwork (the shared convolution parameters and W in our WCNN) to extract invariant feature representation.

TABLE 5
Performance comparison of different parameter settings of WCNN. The trade-off parameters are set such that the value of each term is on the same scale.

β_1	β_2	β_3	λ_i	Rank-1	FAR=0.1%
1	0	0	0.0005	97.3±0.6	94.9±1.1
1	0	0	0.001	97.3±0.4	95.7±0.7
1	0	0	0.002	97.3±0.5	94.6±1.3
1	0.5	0	0.001	97.7±0.5	96.9±0.5
1	1.0	0	0.001	98.4±0.4	97.6±0.3
1	2.0	0	0.001	98.4±0.4	97.5±0.3
1	1.0	0.0005	0.001	98.4±0.3	98.0±0.5
1	1.0	0.001	0.001	98.7±0.3	98.4±0.4
1	1.0	0.002	0.001	98.6±0.3	98.1±0.6

To further investigate the effectiveness of each component of the WCNN, we vary the parameter values in Eq. (14). Note that, to balance the components, we simply set β_1, β_2, β_3 and λ_i to make the value of each component be on the same scale. Then, these parameters are multiplied by two scale factors (0.5 and 2.0). The small scale factor reduces the effectiveness of a component, and the large scale factor enhances the effectiveness of a component. Note that the CASIA NIR-VIS 2.0 database contains 10-fold evaluation experiments and each fold has 214,800 comparisons. According to Table 5, small performance variations are observed under the different settings. As expected, when one or two components in Eq. (14) are dropped, recognition accuracy decreases significantly. These findings indicate that if the parameters of the WCNN are balanced, large parameter ranges occur in which the WCNN can obtain a better result.

5 Conclusions

This paper has developed a Wasserstein CNN approach that uses only one network to project both NIR and VIS images to a compact Euclidean space. The WCNN naturally combines subspace learning and invariant feature extraction into a CNN, and divides the high-level layer of the CNN into two orthogonal subspaces that contain modality-invariant identity information and modality-variant light spectrum information. The Wasserstein distance is used to measure the difference between heterogeneous feature distributions, and it is effective at reducing the sensing gap. To the best of our knowledge, this paper represents the first attempt to formulate a probability distribution learning approach for VIS-NIR matching in the NIR-VIS HFR field. In addition, a low-rank constraint has been studied for alleviating the over-fitting problem on small-scale NIR-VIS face data. An alternating minimization approach has been developed to minimize the joint formulation of the WCNN in an end-to-end manner. Extensive experimental results using three challenging NIR-VIS face recognition databases show that our WCNN methods significantly outperform state-of-the-art NIR-VIS face recognition methods. Experimental results also verify the effectiveness of Wasserstein distance to measure NIR and VIS distributions, resulting in further performance improvements. However, when training samples are

small-scale, the improvement benefited from Wasserstein distance tends to be small. Hence, one future trend is to investigate more distribution measures to accurately match NIR and VIS distributions.

ACKNOWLEDGMENT

We would like to thank the associate editor and the reviewers for their valuable comments and advice. This work is funded by State Key Development Program (Grant No. 2016YFB1001001) and the National Natural Science Foundation of China (Grants No. 61622310 and 61427811).

REFERENCES

[1] R. He, B. C. Lovell, R. Chellappa, A. K. Jain, and Z. Sun, "Editorial: Special issue on ubiquitous biometrics," *Pattern Recognition*, vol. 66, pp. 1–3, 2017.

[2] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C. C. Loy, and X. Wang, "A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution," *Image and Vision Computing*, vol. 56, pp. 28–48, 2016.

[3] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2079–2089, 2017.

[4] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.

[5] X. Tang and X. Wang, "Face photo recognition using sketch," in *IEEE International Conference on Image Processing*, 2002.

[6] Z. Lei, Q. Bai, R. He, and S. Li, "Face shape recovery from a single image using cca mapping between tensor spaces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2019–2030, 2012.

[8] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *IEEE International Conference on Computer Vision*, 2017.

[9] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *IEEE Workshop on Perception Beyond the Visible Spectrum*, 2016, pp. 54–62.

[10] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Z. Li, "Matching NIR face to VIS face using transduction," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 501–514, 2014.

[11] X. Liu, X. Sun, R. He, and T. Tan, "Recent advances on cross-domain face recognition," in *Chinese Conference on Biometric Recognition*, 2016, pp. 147–157.

[12] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 640–652, 2015.

[13] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Conference on Neural Information Processing Systems*, 2014, pp. 1988–1996.

[14] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[16] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *International Conference on Biometrics*, 2016.

[17] S. Saxena and J. Verbeek, "Heterogeneous face recognition with cnns," in *European Conference on Computer Vision Workshops*, 2016, pp. 483–491.

[18] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*, 2012.

[19] D. Yi, Z. Lei, S. Liao, and S. Li, "Shared representation learning for heterogeneous face recognition," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015.

[20] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *International Conference on Machine Learning*, 2013.

[21] J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikainen, "Learning mappings for face synthesis from near infrared to visual light images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 156–163.

[22] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," Beihang University, Beijing, China, Tech. Rep. IRIP-TR-12-FR-001, July 2012.

[23] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen, "Viplfacenet: An open source deep face recognition sdk," *Frontiers of Computer Science*, 2016.

[24] L. Xiao, R. He, Z. Sun, and T. Tan, "Coupled feature selection for cross-sensor iris recognition," in *IEEE conference on Biometrics: Theory, Applications and Systems*, 2013, pp. 1–6.

[25] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 687–694.

[26] R. Wang, J. Yang, D. Yi, and S. Li, "An analysis-by-synthesis method for heterogeneous face biometrics," in *International Conference on Biometrics*, 2009, pp. 319–326.

[27] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[28] X. Gao, J. Zhong, J. Li, and C. Tian, "Face sketch synthesis algorithm based on e-hmm and selective ensemble," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 4, pp. 487–496, 2008.

[29] S. Wang, D. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2216–2223.

[30] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *IEEE International Conference on Computer Vision*, 2013, pp. 2496–2503.

[31] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2015.

[32] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[33] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *International Conference on Biometrics*, 2009, pp. 209–218.

[34] Z. Lei, R. Chu, R. He, S. Liao, and S. Z. Li, "Face recognition by discriminant analysis with gabor tensor representation," in *IAPR/IEEE International Conference on Biometrics*, 2007.

[35] B. Klare and A. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *International Conference on Pattern Recognition*, 2010, pp. 1513–1516.

[36] D. Goswami, C. H. Chan, D. Windridge, and J. Kittler, "Evaluation of face recognition system in heterogeneous environments (visible vs NIR)," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2160–2167.

[37] L. Huang, J. Lu, and Y.-P. Tan, "Learning modality-invariant features for heterogeneous face recognition," in *International Conference on Pattern Recognition*, 2012, pp. 1683–1686.

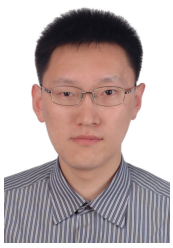
[38] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 639–646, 2011.

[39] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 2, pp. 451–463, 2016.

[40] D. Lin and X. Tang, "Inter-modality face recognition," in *European Conference on Computer Vision*, 2006, pp. 13–26.

[41] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1707–1716, 2012.

[42] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 353–362, 2013.
[43] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *IEEE International Conference on Computer Vision*, 2013, pp. 2088–2095.
[44] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1410–1422, 2013.
[45] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 188–194, 2016.
[46] Z. Li, D. Gong, Q. Li, D. Tao, and X. Li, "Mutual component analysis for heterogeneous face recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, pp. 1–23, 2016.
[47] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Tech. Rep., 2007.
[48] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for nir-vis face recognition," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 2000–2006.
[49] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *International Conference on Machine Learning*, 2017.
[50] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv:1703.10717*, 2017.
[51] B. Su and G. Hua, "Order-preserving wasserstein distance for sequence matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[52] J. Feng, Z. Lin, H. Xu, and S. Yan, "Robust subspace segmentation with block-diagonal prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[53] G. Xavier and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 499–515.
[54] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *CoRR abs/1511.02683*, 2015.
[55] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," *CoRR*, vol. abs/1607.08221, 2016. [Online]. Available: http://arxiv.org/abs/1607.08221
[56] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1123–1128.
[57] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 289–302, 2014.
[58] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, "On effectiveness of histogram of oriented gradient features for visible to near infrared face matching," in *International Conference on Pattern Recognition*, 2014, pp. 1788–1793.
[59] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.



Ran He received the B.E. degree in Computer Science from Dalian University of Technology, the M.S. degree in Computer Science from Dalian University of Technology, and Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences in 2001, 2004 and 2009, respectively. Since September 2010, Dr. He has joined NLPR where he is currently Professor. He currently serves as an associate editor of Neurocomputing (Elsevier) and serves on the program committee of several conferences. His research interests focus on information theoretic learning, pattern recognition, and computer vision.

Xiang Wu received the B.E. degree in Electronic Engineering from University of Science and Technology Beijing in 2013, the M.S. degree in Electronic Engineering from University of Science and Technology Beijing in 2016. He is a research assistant in Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests focus on deep learning, computer vision and biometric recognition.

Zhenan Sun received the BE degree in industrial automation from Dalian University of Technology in 1999, the MS degree in system engineering from Huazhong University of Science and Technology in 2002, and the PhD degree in pattern recognition and intelligent systems from CASIA in 2006. He is a professor in the Institute of Automation, Chinese Academy of Sciences (CASIA). In March 2006, he joined the Center of Biometrics and Security Research (CBSR) in the National Laboratory of Pattern Recognition (NLPR) of CASIA as a faculty member. He is a member of the IEEE and the IEEE Computer Society. His research focuses on biometrics, pattern recognition, and computer vision.

Tieniu Tan received his B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and his M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He was the Director General of the CAS Institute of Automation from 2000-2007, and has been Professor and Director of the NLPR since 1998. He has published more than 300 research papers in refereed journals and conferences in the areas of image processing, computer vision and pattern recognition, and has authored or edited 9 books. He holds more than 30 patents. His current research interests include biometrics, image and video understanding, and information forensics and security.

Dr Tan is a Fellow of CAS, TWAS, BAS, IEEE, IAPR, UK Royal Academy of Engineering, and Past President of IEEE Biometrics Council. He currently serves the President of the Chinese Society of Image and Graphics, Deputy President of the Chinese Association for Artificial Intelligence. He has given invited talks and keynotes at many universities and international conferences, and has received numerous national and international awards and recognitions.