Learning Invariant Deep Representation for NIR-VIS Face Recognition

Ran He, Xiang Wu, Zhenan Sun, Tieniu Tan
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
University of Chinese Academy of Sciences, Beijing 100190, China.
{rhe,znsun,tnt}@nlpr.ia.ac.cn, alfredxiangwu@gmail.com

Abstract

Visual versus near infrared (VIS-NIR) face recognition is still a challenging heterogeneous task due to large appearance difference between VIS and NIR modalities. This paper presents a deep convolutional network approach that uses only one network to map both NIR and VIS images to a compact Euclidean space. The low-level layers of this network are trained only on large-scale VIS data. Each convolutional layer is implemented by the simplest case of maxout operator. The high-level layer is divided into two orthogonal subspaces that contain modality-invariant identity information and modality-variant spectrum information respectively. Our joint formulation leads to an alternating minimization approach for deep representation at the training time and an efficient computation for heterogeneous data at the testing time. Experimental evaluations show that our method achieves **94%** verification rate at FAR=0.1% on the challenging CASIA NIR-VIS 2.0 face recognition dataset. Compared with state-of-the-art methods, it reduces the error rate by **58%** only with a compact **64-D** representation.

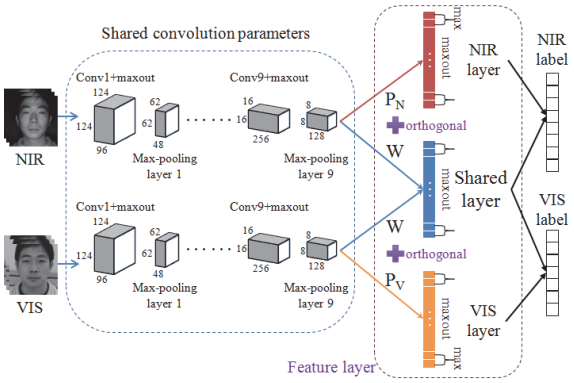

Figure 1: An illustration of our modality invariant deep representation architecture. Orthogonal constraints and maxout operator are used to learn invariant representation and avoid overfitting on a small dataset. At the testing time, both NIR and VIS features are exacted from the shared layer and compared in cosine distance.

1 Introduction

Active near infrared (NIR) imaging technique provides an inexpensive and simple means to enhance the performance of face recognition systems in low light conditions. It has been proved to be less sensitive to visible (VIS) light illumination variations (Zhu et al. 2014), and has been widely used in face identification or authorization applications, such as security surveillance and E-passport. In many real-world applications, face recognition systems almost require individuals to enroll by using their VIS images, which results in the matching problem between NIR and VIS face images. This matching problem is also called the NIR-VIS heterogeneous face recognition problem (Li et al. 2013).

Despite recent advances in the field of deep learning based VIS face recognition (Sun et al. 2014; Taigman et al. 2014; Parkhi, Vedaldi, and Zisserman 2015; Schroff, Kalenichenko, and Philbin 2015), implementing NIR-VIS face recognition efficiently presents serious challenges to current approaches. These challenges may be incurred by two facts. First, since NIR and VIS images are captured from different spectral domains, they have large appearance difference. Hence, the deep convolutional networks trained on VIS data do not contain NIR spectral information so that they fail to deal with NIR images very well. How to utilize large-scale VIS face data to explore modality invariant representation of NIR and VIS face images is an ongoing issue. Second, benefitting from web data, we can easily collect millions of VIS face images. However, pair-wised NIR face images are often unavailable on internet. The collection of large-scale and pair-wised NIR and VIS face images is still expensive. How to apply deep learning on a small NIR-VIS dataset remains a central problem.

Previous NIR-VIS matching methods often use a trick to alleviate the appearance difference problem by removing some principal subspaces that are assumed to contain light spectrum information (Li et al. 2013)(Yi et al. 2015). Observation and results also demonstrate that the appearance of a face is composed of identity information and variation information (e.g., lightings, poses, and expressions) (Chen et al. 2012). Inspired by these observations, this paper presents a deep convolutional network method to learn modality Invariant Deep Representation (IDR) that contains the identity

information of both NIR and VIS face images. Our method employs one single network to map both NIR and VIS images to a compact Euclidean space so that the NIR and VIS images in the embedding space directly correspond to face similarity.

Our convolutional network is first trained on large-scale VIS data. Its convolutional layers and fully connected layer are implemented by the simplest case of maxout operator (Goodfellow et al. 2013). This network makes our learned representation be robust to intra-class variation of individuals. Then, the low-level layers of this network are fixed and fine-tuned to be adaptable to NIR data. The high-level layer is divided into two orthogonal subspaces that contain modality-invariant identity information and modality-variant spectrum information respectively. The orthogonal constraints and the maxout operator in the high-level layer can reduce parameter space and hence avoid the overfitting problem on a mall NIR-VIS dataset. Our joint formulation leads to an alternating minimization approach for deep representation at the training time and an efficient computation for heterogeneous data at the testing time. Extensive experimental evaluations demonstrate that our IDR method learns modality-invariant representation and outperforms state-of-the-art NIR-VIS face recognition methods. Fig. 1 gives an illustration of our IDR method. The main contributions are summarized as follows,

- An effective deep neural network architecture is developed to learn modality invariant representation, and efficiently optimized by alternating minimization. This architecture can naturally combine previous invariant feature extraction and subspace learning into a unified network.
- Two orthogonal subspaces are embedded in the network to model identity and spectrum information respectively. This formulation not only results in one single network structure to extract compact representation but also alleviates the overfitting problem on small-scale data.
- Experimental results on the challenging CASIA NIR-VIS 2.0 face database show that the proposed 64-D IDR advances the best rank-1 accuracy from 86.16% to 95.82% and VR (@FAR=0.001) from 85.80% to 94.03%.

The rest of this paper is organized as follows. We briefly review some related work on heterogeneous biometric recognition in Section 2. In Section 3, we present the details of our IDR method for NIR-VIS face recognition. Section 4 provides experimental results, prior to summary in Section 5.

2 Related Work

During the last decade, heterogeneous biometric recognition has drawn much attention due to the rapid development of various sensors. The notation 'heterogeneous' may refer to NIR vs. VIS (2013; 2015), sketch vs. VIS (2002), 2D vs. 3D (2008), cross-sensor (2013a; 2013b) and different resolutions (2012). Many methods have been proposed to alleviate the appearance difference problem of heterogeneous data. Most of these methods can be generally categorized into three classes: image synthesis, subspace learning and invariant feature extraction (Zhu et al. 2014)(Jin, Lu, and Ruan 2015).

Image synthesis methods aim to synthesize face images from one modality (or domain) into another so that heterogeneous images can be compared in the same distance space. (Wang et al. 2009) applied face analogy to transform a face image from one modality to another. (Wang and Tang 2009) resorted to multiscale Markov random fields to synthesize pseudo-sketch to face photo. Then, (Gao et al. 2008) further used hidden Markov model to learn the nonlinear relationship between face photo and sketch. (Lei et al. 2008) reconstructed a 3D face model from a single 2D face image using canonical correlation analysis (CCA). (Wang et al. 2012), (Huang and Wang 2013) and (Juefei-Xu, Pal, and Savvides 2015) used coupled or joint dictionary learning to reconstruct face images and then performed face recognition.

Subspace learning methods learn mappings to project homogenous data into a common space. CCA and partial least squares (PLS) are two representative methods. (Lin and Tang 2006) proposed a common discriminant feature extraction to incorporate both discriminative and locality information. (Lei et al. 2012) developed a coupled discriminant analysis based on the locality information in kernel space. (Huang et al. 2013) proposed a regularized discriminative spectral regression method to map heterogeneous data into a common spectral space. Recently, (Wang et al. 2013) took feature selection into consideration during common subspace learning. State-of-the-art NIR-VIS results are often obtained by removing some principal subspace components (Yi et al. 2015).

Invariant feature extraction methods try to explore modality-invariant features that are robust to lighting conditions. The current methods are almost based on handcrafted local features, such as local binary patterns (LBP), histograms of oriented gradients (HOG), Difference-of-Gaussian (DoG) and SIFT (Liao et al. 2009)(Klare and Jain 2010)(Goswami et al. 2011). In addition, (Huang, Lu, and Tan 2012) applied sparse representation to learn modality-invariant features. (Zhu et al. 2014) combined Log-DoG filtering, local encoding and uniform feature normalization together to find better feature representation.

Although many efforts have been made, NIR-VIS recognition performance is still quite low compared to VIS recognition performance. For example, the rank-1 accuracy on the challenging CASIA NIR-VIS 2.0 face database is smaller than 90% whereas that on the Labeled Faces in the Wild (LFW) VIS database (Huang et al. 2007) has been more than 99%. The significant improvement in VIS recognition is mainly due to the application of deep learning methods. However, to the best of our knowledge, there are few deep learning works for NIR-VIS face recognition. Hence, it is the time to explore modality invariant deep representation.

3 Invariant Deep Representation

Benefiting from the development of convolutional neural network (CNN), VIS face recognition has made great progress in recent years. This section introduces the idea of subspace decomposition and invariant feature extraction

into CNN to learn modality invariant deep representation for NIR-VIS face recognition.

3.1 Problem Formulation

Let I_V and I_N be the VIS and NIR images respectively. The CNN feature extraction process is denoted as $X_i = Conv(I_i, \Theta_i)$ $(i \in \{N, V\})$, where $Conv()$ is the feature extraction function defined by the ConvNet, X_i is the extracted feature vector, and ϕ_I denotes ConvNet parameters for modality I to be learned. In heterogeneous recognition, one basic assumption is the fact that there is some concept common between heterogeneous samples. Hence, we assume that NIR and VIS face images share some common low-level features. That is, $\Theta_N = \Theta_V = \Theta$ and $X_i = Conv(I_i, \Theta)$. As shown in Fig. 1, the output of the last max-pooling layer represents $X_i \in \mathrm{R}^m$, corresponding to the NIR and VIS channel, respectively. These two channels share the same parameter Θ.

Inspired by the observation that removing spectrum information is helpful for NIR-VIS performance, we further introduce three mapping matrices (i.e., $W, P_i \in \mathrm{R}^{d \times m}$) to model identity invariant information and variant spectrum information. Therefore, the feature representation can be defined as

$$F_i = \begin{bmatrix} F_{\text{shared}} \\ F_{\text{unique}} \end{bmatrix} = \begin{bmatrix} WX_i \\ P_i X_i \end{bmatrix} \ (i \in \{N, V\}) \quad (1)$$

where WX_i and $P_i X_i$ 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 orthogonal constraint on them to make them be unrelated by each other, i.e.,

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

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

$$\begin{aligned} \mathcal{L}(F, c, \Theta, W, P) &= \sum_{i \in \{N,V\}} \text{softmax}(F_i, c, \Theta, W, P_i) \\ &= -\sum_{i \in \{N,V\}} (\sum_{j=1}^{N} \mathbf{1}\{y_{ij} = c\} \log \hat{p}_{ij}) \\ \text{s.t.} \quad & P_i^T W = 0 \ (i \in \{N, V\}) \end{aligned} \quad (3)$$

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 $\mathbf{1}\{\text{a false statement}\} = 0$.

3.2 Optimization Method

Since (3) contains several variables and is non-convex, we develop an alternating minimization method to minimize (3). First, according to the lagrange multipliers, (3) can be reformulated as an unconstrained problem,

$$\begin{aligned} \mathcal{L}(F, c, \Theta, W, P) = &\sum_{i \in \{N,V\}} \text{softmax}(F_i, c, \Theta, W, P_i) \\ &+ \sum_{i \in \{N,V\}} \lambda_i \| P_i^T W \|_F^2 \end{aligned} \quad (4)$$

Algorithm 1: Training the IDR network.

Require: Training set I_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 by (10);
2: **for** $t = 1, \ldots, T$ **do**
3: Fix W, P_i;
4: Update Θ according to back-propagation method;
5: Fix Θ
6: Update W according to (5);
7: Update P_i according to (6);
8: **end for**;
9: **Return** Θ and W;

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

If gradient descent method is used to minimize (4), we should update the parameters W, P_I and Θ. For the CNN parameters Θ, we follow the conventional back-propagation method to update it. The gradients of W and P_I contain two components that can be expressed as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W} = &\sum_{i \in \{N,V\}} \frac{\partial \text{softmax}(F_i, c, \Theta_i, W, P_i)}{\partial W} \\ &+ \sum_{i \in \{N,V\}} \lambda_i P_i P_i^T W \end{aligned} \quad (5)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial P_i} = &\frac{\partial \text{softmax}(F_i, c, \Theta_i, W, P_i)}{\partial P_i} \\ &+ \lambda_i W W^T P_i \end{aligned} \quad (6)$$

Then we update these parameters with a learning rate γ via

$$\Theta^{(t+1)} = \Theta^{(t)} - \gamma \frac{\partial \mathcal{L}}{\partial \Theta^{(t)}} \quad (7)$$

$$W^{(t+1)} = W^{(t)} - \gamma \frac{\partial \mathcal{L}}{\partial W^{(t)}} \quad (8)$$

$$P_i^{(t+1)} = P_i^{(t)} - \gamma \frac{\partial \mathcal{L}}{\partial P_i^{(t)}} \quad (9)$$

Here, we employ the alternating optimization to update all the parameters. As in (Xavier and Bengio 2010), the parameters Θ of CNN is initialized by the pre-trained model and the mapping matrices W, P_I is initialized by

$$W, P_i \sim U\left[-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\right] \quad (10)$$

where $U[-a, a]$ is the uniform distribution in the interval $(-a, a)$ and m is the dimension of original features. The optimization detail is summarized in Algorithm 1.

3.3 Network Architecture

The basic VIS network architecture (the part of share convolution parameters in Fig. 1) and initial values of Θ are trained on a large-scale VIS dataset (Guo et al. 2016). We employ the lightened CNN B network (Wu et al. 2015) as

Method-Dim	100% - EER	Rank-1 accuracy	VF@FAR=1%	VF@FAR=0.1%
DR-64	97.37%±0.21%	90.66%±0.43%	95.77%±0.52%	89.51%±0.47%
IDRm-64	97.98%±0.32%	93.65%±0.85%	97.24%±0.57%	92.56%±1.01%
IDR-64	98.79%±0.11%	95.82%±0.76%	98.58%±0.25%	94.03%±1.06%
DR-128	98.00%±0.20%	93.78%±0.54%	95.05%±0.38%	91.82%±0.79%
IDRm-128	98.30%±0.17%	94.77%±0.33%	97.74%±0.22%	93.45%±0.51%
IDR-128	98.93%±0.17%	97.33%±0.43%	98.89%±0.29%	95.73%±0.76%

Table 1: Equal error rate (EER) (± standard variation), rank-1 accuracy (± standard variation) and verification rate(VF)@false accept rate(FAR) (± standard variation) of the proposed deep learning approach with different settings.

the basic network[1]. The network includes nine convolution layers with four max-pooling layers, followed by the fully connected layer. Softmax is used as the loss function. The training VIS face images are normalized and cropped to 144×144 according to five facial points. To enrich the input data, we randomly cropped the input images into 128×128. The MS-Celeb-1M dataset (Guo et al. 2016), which contains totally 8.5M images for about 100K identities, is employed to train the basic network. Dropout ratio is set to 0.7 for fully connected layer and the learning rate is set to $1e^{-3}$ initially and reduced to $1e^{-5}$ for $4,000,000$ iterations. The trained single model for the basic network obtained 98.90% on the LFW dataset.

Based on the basic VIS network, we develop a modality invariant convolution neural network for NIR-VIS face recognition. The low-level convolution layers are initialized by the pre-trained basic network. We implement two CNN channels with shared parameters to input NIR and VIS images respectively. Then we define the feature layer (as in Fig. 1) that aims to project the low-level features into two orthogonal feature subspaces. In this way, we can leverage the correlated properties of NIR and VIS identities and enforce the domain-specific properties of both modalities. Finally, the softmax loss functions are separately used for NIR and VIS representation as the supervisory signals. Note that since there is a maxout operator in the feature layer, the final feature dimension is $d/2$ when $W \in \mathrm{R}^{d \times m}$. As in VIS training, all NIR and VIS images are cropped and resized to 144×144 pixels and a randomly selected 128×128 regions are fed into the IDR network for NIR-VIS training. The learning rate of the IDR network is set to $1e^{-4}$ initially and reduced to $1e^{-6}$ gradually for around $100,000$ iterations.

4 Experiments

In this section, we perform experiments on the most challenging CASIA NIR-VIS 2.0 face database (Li et al. 2013). We first introduce the database and testing protocols, then present algorithmic analysis and detailed evaluation, as well as comparison with state-of-the-art NIR-VIS methods.

[1]https://github.com/AlfredXiangWu/face_verification_experiment

4.1 Dataset and Protocols

The CASIA NIR-VIS 2.0 Face Database is widely used in NIR-VIS heterogeneous face evaluations because it is the largest public and most challenging NIR-VIS database. Its challenge is due to large variations of the same identity, including lighting, expression, pose, and distance. Wearing glasses or not is also considered to generate Variations. The database is composed of 725 subjects, each with 1-22 VIS and 5-50 NIR images. Each image is randomly gathered so that there are not one-to-one correlations between NIR and VIS images. The database contains two views of evaluation protocols. View 1 is used for super-parameters adjustment, and View 2 is used for training and testing.

For a fair comparison with other results, we follow the standard protocol in View 2. There are 10-fold experiments in View 2. Each fold contains a collection of training and testing lists. Nearly equal numbers of identities are included in the training and testing sets, and are kept disjoint from each other. Training on each fold is many-to-many (i.e., images from NIR and VIS are randomly combined). For each training fold, there are around 2,500 VIS images and around 6,100 NIR images from around 360 subjects. These subjects are mutually exclusive from the 358 subjects in the testing set. That is, the subjects in the training set and testing set are entirely different. The training set in each fold is used for IDR training. For each testing fold, the gallery set always contains a total of 358 subjects, and each subject only has one VIS image. The probe set has over 6,000 NIR images from the same 358 subjects. All the probe set is to be matched against the gallery set, resulting in a similarity matrix of size 358 by around $6,000$.

4.2 Algorithmic Analysis

In this subsection, we give a detailed evaluation of each part of our proposed invariant deep representation (**IDR**) method. We implement two simpler version of IDR. The notation **DR** indicates the IDR without the NIR feature and VIS feature in Fig. 1. That is, we only train one convolutional network without subspace decomposition. Note that there are a large number of parameters in the fully connected layer and the feature layer (i.e., NIR, shared and VIS features in Fig. 1), which result in overfitting on a small-scale NIR-VIS dataset. The maxout operator in the feature layer is

also helpful to alleviate the overfitting problem. Hence, the notation **IDRm** indicates the IDR without maxout operator in the feature layer.

In Table 1, three performance measures are reported for comparison, including equal error rates, rank-1 accuracy and verification rates. We observe that the rank-1 accuracy and VF@FAR=0.1 of DR-128 have been better than the state-of-the-art rank-1 accuracy 86.16% (Yi et al. 2015) and VF@FAR=0.1 85.80% (Juefei-Xu, Pal, and Savvides 2015). DR-128 is trained on a large-scale VIS dataset and fine-tuned on a small-scale NIR-VIS dataset. These improvements indicate that our basic strategy to train a network is efficient for NIR-VIS face recognition problem. In addition, different implementations lead to different recognition results.

We also observe that IDR almost obtains the lowest performance among the three implementations. Particularly, there is a large performance improvement of IDR against DR. Comparing the rank-1 accuracy of IDR and DR, we find that there is nearly 4% rank-1 accuracy difference on the dimensions 64 and 128. These results suggest that the usage of two orthogonal subspaces to learn invariant representation is effective. These orthogonal subspaces can potentially separate light information from identification information so that some hard NIR-VIS pairs are correctly classified.

Comparing IDR with IDRm, we observe that the maxout operator in the last convolutional layer can further reduce equal error rate (EER) and improve verification rates. When feature dimension is 64, EER can be reduced by 40% $(100\% * (2.02 - 1.21)/2.02)$. IDR-128 obtains better results than IDR-64 and achieves the best performance in terms of EER, rank-1 accuracy and verification rate. This indicates that when the number of features is increased, the performance is also increased accordingly. All of these verify the benefit of our IDR and suggest the usage of maxout operator.

4.3 Comparison with State-of-the-art Methods

To verify the performance of IDR, we compare our method with state-of-the-art NIR-VIS recognition methods, including PCA+Sym+HCA (Li et al. 2013), learning coupled feature spaces (LCFS) method (Wang et al. 2013; Jin, Lu, and Ruan 2015), coupled discriminant face descriptor (C-DFD) (Lei, Pietikainen, and Li 2014; Jin, Lu, and Ruan 2015), DSIFT+PCA+LDA (Dhamecha et al. 2014), coupled discriminant feature learning (CDFL) (Jin, Lu, and Ruan 2015), Gabor+RBM+Remove 11PCs (Yi et al. 2015), VIS+NIR reconstruction+UDP (Juefei-Xu, Pal, and Savvides 2015). The results of LCFS, C-DFD and CDFL are from (Jin, Lu, and Ruan 2015), and those of the remaining compared methods are from their own papers. The results of three VIS CNN methods are also discussed, including VGG (Parkhi, Vedaldi, and Zisserman 2015), SeetaFace (Liu et al. 2016) and CenterLoss (Wen et al. 2016).

Fig. 2 plots the receiver operating characteristic (ROC) curves of the proposed method and its three top competitors. For a better illustration, we do not report some ROC curves of other methods if these curves are low. We observe that the methods can be nearly ordered in ascending ROC curves as Gabor+Remove 20PCS (Yi et al. 2015), Gabor+RBM+Remove 11PCs (Yi et al. 2015), VIS+NIR reconstruction+UDP, IDR-64 and IDR-128. Our IDR methods consistently outperform its three competitors when FAR is smaller than 1%. They can significantly improve verification rates especially when FAR is low. This indicates that IDR can correctly classify some difficult NIR-VIS sample pairs. In addition, IDR-128 performs better than IDR-64. This indicates that ROC curves can be further improved if more features are used.

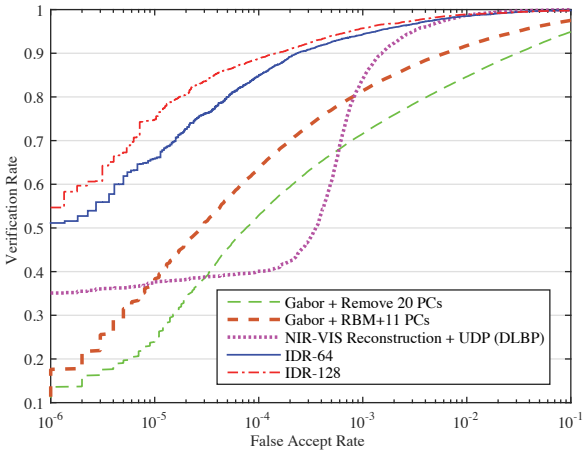

Figure 2: ROC curves of different NIR-VIS face recognition methods.

Table 2 shows the rank-1 accuracy and VR@FAR=0.1% of different NIR-VIS methods. The methods can be ordered in ascending rank-1 accuracy as PCA+Sym+HCA, LCFS, C-DFD, CDFL, DSIFT+PCA+LDA, VIS+NIR reconstruction+UDP, Gabor+RBM+Remove 11PCs, IDR. Except IDR, the best three methods apply several different techniques separately to improve rank-1 accuracy. In contrast, IDR naturally fuses these techniques into a unified neural network framework, which makes IDR achieve the highest accuracy. Our IDR method improves the best rank-1 accuracy from 86.16% to 95.82%. It reduces the error rate by **70%** only with a compact 64-D feature representation. There is also a significant improvement on VR@FAR=0.1%. The improvement is nearly 10% verification rate. As expected, the three CNN methods trained on VIS face data do not work on the NIR-VIS matching problem. They can not further improve verification rates. This is because the large appearance between VIS and NIR domain. All of 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 unique CNN.

5 Conclusion and Future Work

By naturally combining subspace learning and invariant feature extraction into CNNs, this paper has developed an invariant deep representation approach that uses only one network to map both NIR and VIS images to a compact Euclidean space. The low-level layers of this representation are trained on large-scale VIS data. The high-level layer is divided into two orthogonal subspaces that contain modality-

Methods	Rank-1 accuracy	VF@FAR=0.1%	Dimension
PCA+Sym+HCA (2013)	23.70%±1.89%	19.27%	-
LCFS (2013)(2015)	35.4%±2.8%	16.7%	-
C-DFD (2014)(2015)	65.8%±1.6%	46.2	-
DSIFT+PCA+LDA (2014)	73.28%±1.10%	-	-
CDFL (2015)	71.5%±1.4%	55.1%	1000
Gabor+RBM+Remove 11PCs (2015)	86.16%±0.98%	81.29±1.82%	$80 \times 176 = 14080$
VIS+NIR reconstruction+UDP (2015)	78.46%±1.67%	85.80%	$32 \times 32 = 1024$
VGG (2015)	62.09%±1.88%	39.72%±2.85%	4096
SeetaFace (2016)	68.03%±1.66%	58.75%±2.26%	2048
CenterLoss (2016)	87.69%±1.45%	69.72%±2.07%	1024
IDR	**95.82%**±0.76%	**94.03%**± 1.06%	**64**

Table 2: Rank-1 accuracy (± standard variation) and verification rate (± standard variation) at FAR=0.1% on the CASIA 2.0 NIR-VIS face database.

invariant identity information and modality-variant light spectrum information respectively. We have proposed an alternating minimization approach to minimize and the joint formulation of IDR. Experimental results on the challenging CASIA NIR-VIS 2.0 face recognition dataset show that our IDR method significantly outperforms state-of-the-art NIR-VIS face recognition methods.

An intriguing question for future work is whether this IDR framework can be useful for other heterogeneous or cross-modal problems, e.g., cross-sensor iris recognition and sketch-VIS face recognition. We believe that the full potential of IDR is yet to be uncovered in many heterogeneous problems. Another direction is to establish a large-scale NIR-VIS dataset for both training and testing.

Acknowledgments

This work is partially funded by the State Key Development Program (Grant No. 2016YFB1001001), National Natural Science Foundation of China (Grant No. 61473289, 61622310) and Beijing Municipal Science and Technology Project (Grant No. Z141100003714131, Z161100000216144).

References

Biswas, S.; Bowyer, K. W.; and Flynn, P. J. 2012. Multidimensional scaling for matching low-resolution face images. *IEEE TPAMI* 34(10):2019–2030.

Chen, D.; Cao, X.; Wang, L.; Wen, F.; and Sun, J. 2012. Bayesian face revisited: A joint formulation. In *ECCV*.

Dhamecha, T. I.; Sharma, P.; Singh, R.; and Vatsa, M. 2014. On effectiveness of histogram of oriented gradient features for visible to near infrared face matching. In *IEEE ICPR*, 1788–1793.

Gao, X.; Zhong, J.; Li, J.; and Tian, C. 2008. Face sketch synthesis algorithm based on e-hmm and selective ensemble. *IEEE TCSVT* 18(4):487–496.

Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A.; and Bengio, Y. 2013. Maxout networks. In *ICML*.

Goswami, D.; Chan, C. H.; Windridge, D.; and Kittler, J. 2011. Evaluation of face recognition system in heterogeneous environments (visible vs NIR). In *IEEE ICCVW*, 2160–2167.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. *CoRR* abs/1607.08221.

Huang, D.-A., and Wang, Y.-C. F. 2013. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *IEEE ICCV*, 2496–2503.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts.

Huang, X.; Lei, Z.; Fan, M.; Wang, X.; and Li, S. Z. 2013. Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE TIP* 22(1):353–362.

Huang, L.; Lu, J.; and Tan, Y.-P. 2012. Learning modality-invariant features for heterogeneous face recognition. In *ICPR*, 1683–1686.

Jin, Y.; Lu, J.; and Ruan, Q. 2015. Coupled discriminative feature learning for heterogeneous face recognition. *IEEE TIFS* 10(3):640–652.

Juefei-Xu, F.; Pal, D. K.; and Savvides, M. 2015. NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *IEEE CVPR*.

Klare, B., and Jain, A. 2010. Heterogeneous face recogni-

tion: Matching NIR to visible light images. In *ICPR*, 1513–1516.

Lei, Z.; Bai, Q.; He, R.; and Li, S. 2008. Face shape recovery from a single image using cca mapping between tensor spaces. In *IEEE CVPR*.

Lei, Z.; Liao, S.; Jain, A. K.; and Li, S. Z. 2012. Coupled discriminant analysis for heterogeneous face recognition. *IEEE TIFS* 7(6):1707–1716.

Lei, Z.; Pietikainen, M.; and Li, S. Z. 2014. Learning discriminant face descriptor. *IEEE TPAMI* 36(2):289–302.

Li, S. Z.; Yi, D.; Lei, Z.; and Liao, S. 2013. The casia nir-vis 2.0 face database. In *IEEE CVPR Workshops*, 348–353.

Liao, S.; Yi, D.; Lei, Z.; Qin, R.; and Li, S. Z. 2009. Heterogeneous face recognition from local structures of normalized appearance. In *ICB*, 209–218.

Lin, D., and Tang, X. 2006. Inter-modality face recognition. In *IEEE ECCV*, 13–26.

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

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition. In *BMVC*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE CVPR*.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *NIPS*, 1988–1996.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, 1701–1708.

Tang, X., and Wang, X. 2002. Face photo recognition using sketch. In *IEEE ICIP*.

Wang, X., and Tang, X. 2009. Face photo-sketch synthesis and recognition. *IEEE TPAMI* 31(11):1955–1967.

Wang, R.; Yang, J.; Yi, D.; and Li, S. 2009. An analysis-by-synthesis method for heterogeneous face biometrics. In *ICB*, 319–326.

Wang, S.; Zhang, D.; Liang, Y.; and Pan, Q. 2012. Semi-coupled dictionary learning with applications to image super-resolution and photosketch synthesis. In *IEEE CVPR*, 2216–2223.

Wang, K.; He, R.; Wang, W.; Wang, L.; and Tan, T. 2013. Learning coupled feature spaces for cross-modal matching. In *ICCV*, 2088–2095.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2015. A light CNN for deep face representation with noisy labels. *CoRR abs/1511.02683*.

Xavier, G., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of AISTATS*, 499–515.

Xiao, L.; He, R.; Sun, Z.; and Tan, T. 2013a. Coupled feature selection for cross-sensor iris recognition. In *BTAS*, 1–6.

Xiao, L.; Sun, Z.; He, R.; and Tan, T. 2013b. Margin based feature selection for cross-sensor iris recognition via linear programming. In *ACPR*, 246–250.

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

Zhu, J.-Y.; Zheng, W.-S.; Lai, J.-H.; and Li, S. Z. 2014. Matching NIR face to VIS face using transduction. *IEEE TIFS* 9(3):501–514.