Disentangled Variational Representation for Heterogeneous Face Recognition

Xiang Wu,[1,2] **Huaibo Huang,**[1,2,3] **Vishal M. Patel,**[4] **Ran He,**[1,2*] **Zhenan Sun**[1,2]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC), CASIA, Beijing, China
[2]National Laboratory of Pattern Recognition (NLPR), CASIA, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4]Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA
alfredxiangwu@gmail.com, huaibo.huang@cripac.ia.ac.cn,
vpatel36@jhu.edu, {rhe, znsun}@nlpr.ia.ac.cn

Abstract

Visible (VIS) to near infrared (NIR) face matching is a challenging problem due to the significant domain discrepancy between the domains and a lack of sufficient data for training cross-modal matching algorithms. Existing approaches attempt to tackle this problem by either synthesizing visible faces from NIR faces, extracting domain-invariant features from these modalities, or projecting heterogeneous data onto a common latent space for cross-modal matching. In this paper, we take a different approach in which we make use of the Disentangled Variational Representation (DVR) for cross-modal matching. First, we model a face representation with an intrinsic identity information and its within-person variations. By exploring the disentangled latent variable space, a variational lower bound is employed to optimize the approximate posterior for NIR and VIS representations. Second, aiming at obtaining more compact and discriminative disentangled latent space, we impose a minimization of the identity information for the same subject and a relaxed correlation alignment constraint between the NIR and VIS modality variations. An alternative optimization scheme is proposed for the disentangled variational representation part and the heterogeneous face recognition network part. The mutual promotion between these two parts effectively reduces the NIR and VIS domain discrepancy and alleviates over-fitting. Extensive experiments on three challenging NIR-VIS heterogeneous face recognition databases demonstrate that the proposed method achieves significant improvements over the state-of-the-art methods.

Introduction

In recent years, methods based on deep convolution neural network (CNN) have shown impressive performance improvements for face detection and recognition problems (Parkhi, Vedaldi, and Zisserman 2015; Wu et al. 2018a). Despite the success of CNN-based methods in addressing various challenges in face recognition such as variations in pose, expression, aging, occlusion, disguise, and illumination, they are specifically designed to recognize face images that are collected at or near the visible (VIS) domain. However, in many real-world applications such as surveillance at night-time and in low-light conditions, one has to be able to recognize faces collected in thermal or near infrared (NIR) domains. The performance of many CNN-based face recognition methods often degrades significantly when confronted by the NIR face images. This is mainly due to the significant distributional change between the NIR and VIS domains.

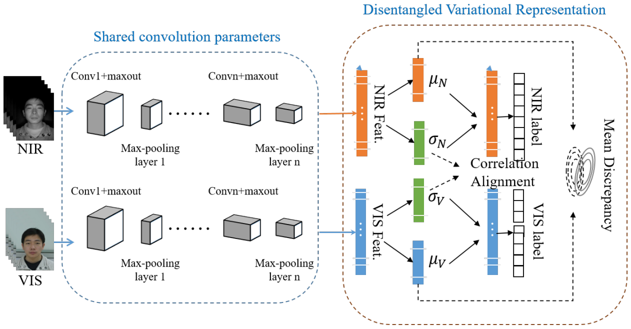

Figure 1: An overview of the proposed Disentangled Variational Representation (DVR) approach for VIS-NIR matching. The NIR and VIS representations x_N and x_V are disentangled into (μ_N, σ_N) and (μ_V, σ_V), respectively. We assume that there is a linear relationship, P, between lighting variations, i.e., $\sigma_V = P\sigma_N$. The mean discrepancy is used to measure the difference between the NIR and VIS distributions in the latent space. The reconstructions $\hat{x}_N$ and $\hat{x}_V$ are obtained from the likelihood $p(x_N|z_N)$ and $p(x_V|z_V)$, respectively and are constrained by the cross-entropy loss.

Another issue that one has to overcome when designing CNN-based models for heterogeneous face recognition (HFR) is over-fitting, which happens due to the lack of sufficient training samples. One of the reasons why CNN-based face recognition methods provide impressive performance improvements on various face recognition benchmarks is that they are trained on thousands and millions of annotated face images often downloaded from the internet. In contrast, there is no publicly available large-scale annotated NIR face dataset for training deep networks. As a result, CNNs trained on small-scale NIR data often tend to overfit. Hence, it is necessary to explore other methods that can deal with this issue in HFR.

Various methods have been developed in the literature for VIS to NIR cross-modal face recognition (Li et al. 2013;

[*]corresponding author

Reale et al. 2016). In particular, methods such as (Liu et al. 2016; He et al. 2017; Zhang et al. 2017; Wu et al. 2018b; Song et al. 2018; He et al. 2018; Di, Zhang, and Patel 2018) attempt to reduce the domain gap between the NIR and VIS domains and learn domain invariant representations for HFR.

In contrast, we take a different approach in which we make use of the Disentangled Variational Representation (DVR) to deal with the two aforementioned challenges. First, inspired by the observation that the facial appearance is composed of the identity information and the variation information, as shown in Fig. 1, we assume that there exists an independent latent variable, which can be composed of an intrinsic variable for identity and an intra-personal variable for within-person variation. Second, benefiting from the variational lower bound (Kingma and Welling 2014) to tackle the marginal likelihood estimation, we model the approximate posterior and obtain disentangled latent variable. Next, when imposing the minimization of the identity information for the same subject and the assumption of correlation alignment (Sun and Saenko 2016) between different modality variations, we obtain more compact and discriminative disentangled latent space for DVR. Although there are large light spectrum variations, spectrum variations are often assumed to be on linear subspaces. Hence, we employ a relaxed correlation alignment item to constrain the variations of different modalities. Furthermore, generating samples from the approximate posterior significantly alleviates the need for having large number of samples during training the fully connected layers of deep HFR models. Since the effectiveness of the generated samples from the likelihood depends on the estimated approximate posterior, we propose an alternative optimization approach for the DVR framework during training in which HFR network can contribute to the disentangled representation training and vice versa.

To summarize, the following are our main contributions:

- An end-to-end DVR framework is developed for cross-modal NIR-VIS face matching. We introduce a variational lower bound to estimate the posterior and optimize the latent variable space, aiming at disentangling the NIR and VIS face representations.
- We propose to minimize the identity information for the same subject and the relaxed correlation alignment constraint on modality variations that facilitate modeling the compact and discriminative disentangled latent variable spaces for heterogeneous modalities.
- An alternative optimization is proposed to provide mutual promotion between HFR network and disentangled variational representation part. Thus, DVR can both reduce the domain discrepancy and alleviate over-fitting.
- Extensive experimental results are conducted on three HFR databases, including the CASIA NIR-VIS 2.0 database (Li et al. 2013), the Oulu-CASIA NIR-VIS database (Chen et al. 2009) and the BUAA-VisNir database (Huang, Sun, and Wang 2012), and comparisons are performed against several recent state-of-the-art approaches. Furthermore, an ablation study is conducted to demonstrate the improvements obtained by various components of the proposed method.

Related Work

We follow the notations in (Zhu et al. 2014; He et al. 2017; 2018; Song et al. 2018) while providing a brief survey of HFR and disentangled representation learning.

Heterogeneous Face Recognition (HFR)

The problem of HFR has gained a lot of traction in recent years (Xiao et al. 2013; Ouyang et al. 2016). According to (Zhu et al. 2014), the existing methods are divided into the following three main categories:

Latent subspace learning aims to project the heterogenous data onto a common latent space in which the relevance of heterogeneous data can be measured. Lin (Lin and Tang 2006) proposed a Common Discriminant Feature Extraction (CDFE) method to incorporate both discriminative and locality information. By introducing feature selection via nuclear norm, a common subspace learning was employed in (Wang et al. 2016). Shao *et al* (Shao, Kit, and Fu 2014) project NIR and VIS data into a generalized subspace where each NIR sample can be represented by a combination of VIS samples. Restricted Boltzmann Machines (RBMs) are employed in (Yi et al. 2015) to learn a shared representation between different domains and then Principal Component Analysis (PCA) is applied to remove the redundancy and heterogeneity. Wang *et al* (Wang et al. 2015) propose several deep neural network-based methods with Canonical Correlation Analysis (CCA) in unsupervised subspace feature learning for HFR. He *et al* (He et al. 2017; 2018) divide the high-level representation into two orthogonal subspaces to obtain domain-invariant identity information and domain-related spectrum information.

Modality-invariant feature learning explores domain-invariant features that are only related to the face identity. Traditional methods are based on the handcrafted local features (Liao et al. 2009; Klare, Li, and Jain 2011; Goswami et al. 2011), including Local Binary Patterns (LBP), Gabor features (Lei et al. 2007), Histograms of Oriented Gradients (HOG) and Difference of Gaussian (DoG). Liao *et al* (Liao et al. 2009) combine DoG filtering and multi-block LBP to encode NIR and VIS images. Klare *et al* (Klare, Li, and Jain 2011) utilize HOG features with sparse representation to improve the performance of HFR. Goswami *et al* (Goswami et al. 2011) combine the LBP histogram representation with Linear Discriminant Analysis (LDA) to extract domain invariant features. As for deep learning, Kan *et al* (Kan, Shan, and Chen 2016) address the discriminant domain invariant feature learning by analyzing the within-class and between-class scatter. Coupled Deep Learning (CDL) (Wu et al. 2018b) utilizes nuclear norm constraint on fully connected layer to alleviate overfitting, and proposes a cross-modal ranking to reduce domain discrepancy. He *et al* (He et al. 2018) decrease the domain gap by Wasserstein distance to obtain domain invariant features for HFR.

Data synthesis attempts to address the domain discrepancy at image level by transforming face images from one modality into another via image synthesis. Data synthesis is first proposed to synthesize and recognize a sketch image from a face photo in (Tang and Wang 2003). Wang (Wang and Tang 2009) applies Markov Random Field (MRF) to transform pseudo-sketch to face photo in a multi-scale way. In (Juefei-Xu, Pal, and Savvides 2015), joint dictionary learning is used to reconstruct face images and then perform face matching. Lezama *et al* (Lezama, Qiu, and Sapiro 2017) propose a cross-spectral hallucination and low-rank embedding to synthesize a heterogeneous image in a patch way. With developments of a photo-realistic synthesis image by Generative Adversarial Network (GAN) (Goodfellow et al. 2014), "recognition via generation" (Zhao et al. 2017; Huang et al. 2017; Hu et al. 2018) is drawn attention by lots of researchers. Song *et al* (Song et al. 2018) utilize a Cycle-GAN (Zhu et al. 2017) to realize a cross-spectral face hallucination, facilitating heterogeneous face recognition via generation. However, due to the small number of images in the training set, there are still challenges to synthesize photo-realistic VIS face images from NIR images.

Learning to Disentangled Representations

Early work (Schmidhuber 1992) attempts to disentangle representations in an autoencoder via penalizing predictability of latent variables. A variant of Boltzmann Machine (Desjardins, Courville, and Bengio 2012) is used to disentangle factors of variations in the training data. Kingma (Kingma and Welling 2014) propose the Variational Auto-Encoder (VAE) framework to achieve limited disentangling performance on simple datasets. Matthey *et al* (Matthey et al. 2017) augment the original VAE framework with a single hyper-parameter β, called β-VAE, that controls the degree of disentanglement in the latent representations. Besides, FactorVAE (Kim and Mnih 2018) is proposed to disentangle by encouraging the distribution of representation to be factorial and independent across the dimensions. Chen (Chen et al. 2012) propose joint Bayesian formulation to decompose a face representation into three parts, including intrinsic difference, transformation difference and noise. An expectation maximization-like learning procedure is employed to optimize the joint formulation and they achieve promising performance on the face recognition tasks. Shi *et al* (Shi et al. 2017) extend the original joint Bayesian approach by modeling the gallery and probe images using two different Gaussian distributions to propose a heterogeneous joint Bayesian approach for HFR.

Proposed Method

We begin this section by reviewing the Wasserstein CNN method (He et al. 2018) that introduces a probabilistic framework for HFR and shows promising results. Based on the Wasserstein CNN, we give the details of our disentangled variational representation method and the corresponding optimization scheme.

Revisiting Wasserstein CNN

Let $x_N \in \mathbb{R}^d$ and $x_V \in \mathbb{R}^d$ denote the NIR and VIS domain data representations, respectively. In Wasserstein CNN (He et al. 2018), it is assumed that the data distributions of the representations for the same identity follow a Gaussian distribution. Hence, $x_N \sim \mathcal{N}(m_N, C_N)$ and $x_V \sim \mathcal{N}(m_V, C_V)$, where m_N, m_V are the mean vectors and C_N, C_V are the covariance matrices. The 2-Wasserstein distance between x_N and x_V corresponding to the same identity is defined as

$$\begin{aligned} W(x_N, x_V) = {} & \|m_N - m_V\|_2^2 \\ & + \mathrm{trace}(C_N + C_V - 2(C_V^{\frac{1}{2}} C_N C_V^{\frac{1}{2}})). \end{aligned} \tag{1}$$

Due to the ability of measuring the consistency between two distributions, Eq. (1) is used to reduce the domain gap between the NIR and VIS images. However, the Wasserstein distance is directly imposed on the NIR and VIS representations, which are obtained from a CNN. It is well-known that CNN-based NIR and VIS representations contain various high-level information including identity, spectrum, pose, noise, etc., which are not disentangled. Therefore, directly matching representation distributions may not lead to better performance especially when the training set is not large enough for HFR.

Disentangled Variational Representation

Let $\{x^{(i)} \in \mathbb{R}^d\}_{i=1}^N$ and $\{z^{(i)} \in \mathbb{R}^h\}_{i=1}^N$ denote N observations and the independent latent variables corresponding to one identity, respectively. For each sample $x^{(i)}$, we can obtain

$$z^{(i)} = \mu^{(i)} + \epsilon \odot \sigma^{(i)}, \tag{2}$$

where $\mu^{(i)}$ represents the identity information, $\sigma^{(i)}$ contains variations, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$, $\mu^{(i)}, \sigma^{(i)}, \epsilon \in \mathbb{R}^d$, and $\odot$ denotes the Hadamard product. Note that the marginal likelihood $p(x) = \int p(x|z)p(z)$ is intractable. Hence, different from the common simplifying assumptions about the marginal or posterior probabilities, we introduce the variational lower bound (or evidence lower bound, ELBO)

$$\begin{aligned} \log p(x^{(i)}) \geq {} & -\mathrm{KL}(q(z|x^{(i)})||p(z)) \\ & + \mathbb{E}_{q(z|x^{(i)})}\left[\log p(x^{(i)}|z)\right], \end{aligned} \tag{3}$$

where $q_\phi(z|x^{(i)})$ can be implemented by a probabilistic encoder, $q_\phi(z|x^{(i)}) \sim \mathcal{N}(z; \mu^{(i)}, \sigma^{2(i)}\mathrm{I})$, and ϕ denotes the parameters. Note that the posterior $p(x^{(i)}|z)$ can be treated as the reconstruction part. Let the prior over the latent variables z be a centered isotropic multivariate Gaussian $p(z) \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$. As a result, the disentangled formulation in Eq. (3) can be treated as a variational autoencoder (Kingma and Welling 2014).

Let $z_N \in \mathbb{R}^h, z_V \in \mathbb{R}^h$ represent the latent variables corresponding to the NIR and VIS representations $x_N \in \mathbb{R}^d, x_V \in \mathbb{R}^d$, respectively. Then, one can approximate the posterior as follows:

$$\begin{aligned} q_N(z_N|x_N^{(i)}) \sim \mathcal{N}(z_N; \mu_N^{(i)}, \sigma_N^{2(i)}\mathrm{I}) \\ q_V(z_V|x_V^{(i)}) \sim \mathcal{N}(z_V; \mu_V^{(i)}, \sigma_V^{2(i)}\mathrm{I}), \end{aligned} \tag{4}$$

where $z_N = \mu_N + \epsilon \odot \sigma_N$, $z_V = \mu_V + \epsilon \odot \sigma_V$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$. Here, ϕ_N and ϕ_V denote the parameters of the NIR and VIS approximate posterior estimator, respectively. In NIR-to-VIS face recognition, the main discrepancy comes from the variation in the light spectrum of NIR and VIS domains. We assume that the light spectrum variations are related as follows

$$\sigma_V = P\sigma_N, \tag{5}$$

where $P \in \mathbb{R}^{h\times h}$ is a correlation alignment matrix. Different from (Sun and Saenko 2016), we assume that there is a linear relationship between covariance matrices rather than requiring them to be similar. Since the latent variables z_N and z_V are independent, we impose an orthogonality constraint on P. Therefore, Eq. (4) can be reformulated as

$$\begin{aligned} q_N(z_N|x_N^{(i)}) &\sim \mathcal{N}(z_N; \mu_N^{(i)}, \sigma_N^{2(i)}\mathrm{I}) \\ q_V(z_V|x_V^{(i)}) &\sim \mathcal{N}(z_V; \mu_V^{(i)}, \sigma_V^{2(i)}\mathrm{I}) \\ \text{s.t. } \sigma_V = P\sigma_N,& \quad P^\top P = I. \end{aligned} \tag{6}$$

The correlation alignment matrix P plays the role of constraining the variations of σ_N and σ_V. It makes the representations of NIR and VIS images vary in a subspace. Experimental results also verify the effectiveness of this correlation alignment constraint. Furthermore, since μ_N and μ_V represent the identity information, benefiting from the Wasserstein CNN, we minimize $\|\mu_N - \mu_V\|_2^2$ for the same identities to reduce the domain discrepancy.

With the above definitions, the proposed DVR formulation is as follows

$$\begin{aligned} \mathcal{J}_{\text{DVR}} = \; & -\underbrace{\Big[\mathrm{KL}(q(z_N|x_N^{(i)})||p(z_N)) + \mathrm{KL}(q(z_V|x_V^{(i)})||p(z_V))\Big]}_{\text{approximate posterior estimator parts}} \\ & + \underbrace{\mathbb{E}\Big[\log p(x_N^{(i)}|z_N)\Big] + \mathbb{E}\Big[\log p(x_V^{(i)}|z_V)\Big]}_{\text{reconstruction parts}} \\ & + \underbrace{\lambda_1\|\mu_N^{(i)} - \mu_V^{(i)}\|_2^2}_{\text{mean discrepancy part}} \\ & \text{s.t. } \sigma_V = P\sigma_N, \quad P^\top P = I. \end{aligned} \tag{7}$$

Using the Lagrange multipliers and the reparameterization trick (Kingma and Welling 2014), Eq. (7) can be reformulated as

$$\begin{aligned} \mathcal{J}_{\text{DVR}} = \; & \underbrace{-\tfrac{1}{2}\sum_j \Big(1 + \log\sigma_{Nj}^{2(i)} - \mu_{Nj}^{2(i)} - \sigma_{Nj}^{2(i)}\Big)}_{\text{NIR approximate posterior estimator}} \\ & \underbrace{-\tfrac{1}{2}\sum_j \Big(1 + \log\sigma_{Vj}^{2(i)} - \mu_{Vj}^{2(i)} - \sigma_{Vj}^{2(i)}\Big)}_{\text{VIS approximate posterior estimator}} \\ & \underbrace{+\mathbb{E}\Big[\log p(x_N^{(i)}|z_N)\Big] + \mathbb{E}\Big[\log p(x_V^{(i)}|z_V)\Big]}_{\text{reconstruction parts}} \\ & \underbrace{+\lambda_1\|\mu_N^{(i)} - \mu_V^{(i)}\|_2^2}_{\text{mean discrepancy part}} \\ & \underbrace{+\lambda_2\|\sigma_V - P\sigma_N\|_2^2 + \lambda_3\|P^\top P - I\|_F^2}_{\text{correlation alignment constraint}}, \end{aligned} \tag{8}$$

where j denotes the j-th element of vectors $\mu_N^{(i)}, \mu_V^{(i)}, \sigma_N^{(i)}$ and $\sigma_V^{(i)}$, $\|\cdot\|_F^2$ denotes the Frobenius norm, and λ_1, λ_2 are the trade-off parameters.

As for the reconstruction parts, we let $p(x|z)$ be a multivariate Gaussian which is computed from z with a multilayer perceptron (MLP). Therefore, given $x_N^{(i)}$ and $x_V^{(i)}$ with the label y, we can generate $\hat{x}_N^{(i)}$ and $\hat{x}_V^{(i)}$ from the likelihood $p(x_N^{(i)}|z_N)$ and $p(x_V^{(i)}|z_V)$, respectively. In (Kingma and Welling 2014), the L2 losses $\|x_N^{(i)} - \hat{x}_N^{(i)}\|_2^2$ and $\|x_V^{(i)} - \hat{x}_V^{(i)}\|_2^2$ are used for the reconstruction parts. In DVR, except for the L2 reconstruction loss, we further impose the cross-entropy loss between the reconstructions $\hat{x}_N^{(i)}$ and $\hat{x}_V^{(i)}$ sampled from the posteriors and the identity label y.

Heterogeneous Recognition Network

Given NIR and VIS face images, I_N and I_V respectively, we denote the CNN features as $x_i = f(I_i; \Theta), i \in \{N, V\}$. The output of a CNN feature is normally fed into a softmax layer for supervised training,

$$\mathcal{J}_{\text{cls}} = \text{softmax}(x_i; W, \Theta), i \in \{N, V\}. \tag{9}$$

Given a training sample $(x_i, y), i \in \{N, V\}$, we can generate $(\hat{x}_i, y), i \in \{N, V\}$ using the DVR framework, which can also be fed into a softmax layer as follows

$$\mathcal{J}_{\text{cls}} = \text{softmax}(x_i, y; W, \Theta) + \text{softmax}(\hat{x}_i, y; W, \Theta), i \in \{N, V\}\;. \tag{10}$$

On the one hand, benefiting from the generated samples $\hat{x}_N$ and $\hat{x}_V$, the CNN feature extraction part $f(\cdot; \Theta)$ can be better optimized and more robust, especially when the training sets for HFR are not large enough. On the other hand, the more robust the CNN feature extraction $f(\cdot; \Theta)$ is, the more precisely the approximate posteriors $q(z_i|x_i), i \in \{N, V\}$ in DVR can be estimated. Inspired by this assumption, we propose an alternative optimization method to obtain domain-invariant representations for HFR.

Optimization

In this section, we present an alternative optimization method for the DVR framework. The CNN feature extraction part $f(\cdot; \Theta)$ is initialized by a pre-trained model. First, we directly optimize the approximate posteriors $q(z_i|x_i), i \in \{N, V\}$ until convergence by Eq. (8), but without mean discrepancy and correlation alignment parts, from the random initialization. Second, we generate $\hat{x}_N$ and $\hat{x}_V$ according to Eq. (2) and Eq. (4). We then fix the parameters ϕ_N and ϕ_V in approximate posterior estimator parts and compute Eq. (10) as the loss function to optimize the parameters of the recognition network Θ and W. Finally, the parameters Θ and W in the recognition network are fixed. We utilize the output $x_i = f(I_i; \Theta), i \in \{N, V\}$ as the input, which contributes to the optimization of the approximate posterior estimator parts. The optimization details are summarized in Algorithm 1.

Regarding testing for HFR, we directly employ the outputs of heterogeneous recognition network $f(\cdot; \Theta)$ to obtain

Algorithm 1 Disentangled Variational Representation (DVR) Training.

```
Require: Training set: NIR images $I_N$, VIS images $I_V$, the
    learning rate $\alpha$ and the trade-off parameters $\lambda_1, \lambda_2, \lambda_3$.
Ensure: The CNN parameters $\Theta, W$, the approximate posterior estimators $\phi_N, \phi_V$ and correlation alignment matrix $P$.
Initialize $\Theta, W$ by pre-trained model;
Obtain $x_N = f(I_N; \Theta), x_V = f(I_V; \Theta)$;
Initialize $\phi_N, \phi_V, P$ randomly;
for $t = 1, \ldots, T$ do
    Optimize $\phi_N, \phi_V$ without mean discrepancy and correlation alignment parts;
end for;
for $t = 1, \ldots, T$ do
    Given $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$, generate $\hat{x}_N$ and $\hat{x}_V$ via Eq. (2) and Eq. (4);
    Compute loss $\mathcal{J}_{\text{cls}}$ via Eq. (10)
    Fix $\phi_N, \phi_V, P$;
        Update $\Theta, W$ via back-propagation;
    Obtain $x_N = f(I_N; \Theta), x_V = f(I_V; \Theta)$;
    Fix $\Theta, W, P$
        Update $\phi_N, \phi_V$ by Eq. (8);
    Fix $\Theta, W, \phi_N, \phi_V$
        Update $P$ by gradient descent;
end for;
Return $\Theta, W, \phi_N, \phi_V, P$;
```

$x_i (i \in \{N, V\})$ as feature representations. The cosine distance is used to compute the similarity score between different heterogenous representations for evaluations. Note that the parameters ϕ_N, ϕ_V, P of disentangled variational part and correlation alignment part are not utilized for testing. These two parts aim to disentangle representations and play a role of regularization; therefore, they are only utilized for training to reduce the domain discrepancy and alleviate overfitting. Experimental results demonstrate that these two parts can facilitate convolutional layers to learn a better feature representation.

Experimental Results

In this section, the proposed variational representation learning framework is systemically evaluated against several state-of-the-art HFR methods. We follow the experimental settings proposed in (He et al. 2017)(Wu et al. 2018b)(Song et al. 2018) and mainly employ NIR and VIS images to perform experiments. Both quantitative results and qualitative results are reported.

Datasets and Protocols

Three publicly available VIS-to-NIR face recognition datasets are used to evaluate the performance of different HFR methods.

The CASIA NIR-VIS 2.0 Face Database (Li et al. 2013) is the largest and most challenging NIR-VIS heterogeneous face recognition database due to the large variations in lighting, expression and pose. It consists of 725 identities, each with 1 to 22 VIS and 5 to 50 NIR images. It consists of 10-fold experiments. For training, there are about 2,500 VIS and 6,100 NIR images from 360 identities. For testing, the gallery set in each fold is constructed from 358 identities and each identity only has one VIS image. The probe set contains over 6,000 NIR images from the same 358 identities. All the NIR images in the probe set are to be matched against the VIS images in the gallery set, resulting in a 6000×358 similarity matrix. The Rank-1 accuracy and verification rate (VR)@ false accept rate (FAR)=0.1% are reported.

The Oulu-CASIA NIR-VIS Database (Chen et al. 2009) contains 80 identities with 6 expression variations. Following the protocols in (He et al. 2018), we select 20 identities as the training set and 20 identities as the testing set. Eight face images from each expression are randomly selected from both NIR and VIS sets. Hence, there are totally 96 images per each subject. All the VIS images of the 20 subjects are used as the gallery set and all the NIR images are treated as the probe set. The similarity matrix between the probe set and the gallery set is of size 960 × 960. The rank-1 accuracy, VR@FAR=1% and VR@FAR=0.1% are reported for comparisons.

The BUAA-VisNir Face Database (Huang, Sun, and Wang 2012) consists of data from 150 subjects with 9 VIS and 9 NIR face images per subject. The training set and testing set are composed of 900 images from 50 identities and 1800 images from the remaining 100 identities, respectively. Only one VIS image is selected in the gallery set and the probe set contains 900 NIR images during testing. The similarity matrix between the probe set and the gallery set is of size 900 × 100. The rank-1 accuracy, VR@FAR=1% and VR@FAR=0.1% are reported for comparisons.

Implementation Details

We employ the Light CNN (Wu et al. 2018a) as a basic network architecture for HFR. Both LightCNN-9 and LightCNN-29 models[1] are used as the backbone networks, which are pre-trained on the MS-Celeb-1M dataset (Guo et al. 2016). All the images in the training set are aligned to 144×144 and randomly cropped to 128×128 as the input. Stochastic gradient descent (SGD) is used, where the momentum is set to 0.9 and weight decay is set to $5e$-4. The learning rate is set to $1e$-4 initially and reduced to $5e$-5 gradually. The batch size is set to 128 and the dropout ratio is 0.5.

A multilayer perceptron (MLP) is used to model the DVR parts. It contains four hidden layers with h dimensions to represent μ_N, μ_V, σ_N and σ_V. Moreover, the correlation alignment matrix P is an $h \times h$ matrix. Specifically, in the experiments, the dimension h is set equal to 64. The input and the output layers are both 256-d, which are similar to the dimensions of features from the face recognition network. During training, the parameters of MLP are initialized by a Gaussian, while P is initialized by an identity matrix I. Adam (Kingma and Ba 2015) is used for back-propagation and the initial learning rate is set $1e$-3 and gradually reduced to $1e$-5. The batch size is set to 128. The trade-off parame-

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

Table 1: The ablation study for DVR. Both LightCNN-9 and LightCNN-29 are used as the backbones.

Backbone	Disentangled Variational Part	Mean Discrepancy	Correlation Alignment	CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS			BUAA-VisNir		
				Rank-1	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%
LightCNN-9	-	-	-	97.1	93.7	93.8	80.4	43.8	94.8	94.3	83.5
	√	-	-	98.0	97.3	96.3	85.9	50.7	96.5	95.8	88.3
	√	√	-	98.2	98.1	98.0	88.6	61.3	97.3	96.6	91.0
	√	√	√	**99.1**	**98.6**	**99.3**	**89.7**	**65.8**	**97.9**	**97.0**	**92.8**
LightCNN-29	-	-	-	98.1	97.4	99.0	93.1	68.3	96.8	97.0	89.4
	√	-	-	99.0	99.1	**100.0**	95.2	79.8	98.0	97.9	93.0
	√	√	-	99.5	99.3	**100.0**	96.5	83.0	98.9	98.4	95.6
	√	√	√	**99.7**	**99.6**	**100.0**	**97.2**	**84.9**	**99.2**	**98.5**	**96.9**

ters λ_1, λ_2 and λ_3 are set equal to 1.0, 0.1 and 0.001, respectively.

Analysis of the Proposed Method

We first compare the performance between LightCNN-9 and LightCNN-29 models since they are used as the backbones for HFR. As shown in Table 1, the LightCNN-29 achieves better performance than LightCNN-9 on all the three HFR databases. This clearly shows that LightCNN-29 is more suitable and robust as a backbone network for HFR.

The aim of the proposed DVR is to model the disentangled latent variables z_N and z_V via $q(z_N|x_N)$ and $q(z_V|x_V)$ for NIR and VIS representations x_N and x_V, respectively. And then, we can easily sample $\hat{x}_N$ and $\hat{x}_V$ according to the likelihood $p(x_N|z_N)$ and $p(x_V|z_V)$. Table 1 presents that on the Oulu-CAISA NIR-VIS database, with the disentangled variational part, the performance on VR@FAR=0.1% is improved from 43.8% to 50.7% for LightCNN-9 and from 68.3% to 79.8% for LigthCNN-29, respectively. The results indicate that DVR can alleviate the lack of training data for HFR.

Similar with the Wasserstein CNN, minimizing mean discrepancy on μ_N and μ_V can significantly reduce the domain gap, which achieves $\mathbf{0.8}\%$, $\mathbf{10.6}\%$ and $\mathbf{2.7}\%$ improvements on VR@FAR=0.1% with LightCNN-9 for CASIA NIR-VIS 2.0, Oulu-CASIA NIR-VIS and BUAA-VisNir, respectively. Furthermore, imposing the correlation alignment constraint in Eq. (5) can also boost the performance, which indicates that the assumptions of modeling the light spectrum variations via correlation alignment is reasonable and effective.

As shown in Table 1, the improvements benefiting from three parts, including disentangled variational part, mean discrepancy part and correlation alignment constraint, verifies that our DVR method can significantly reduce the domain discrepancy and alleviate overfitting even if the number of training samples is not large enough.

Comparisons

The performance of the proposed DVR method based on both LightCNN-9 and LightCNN-29 is compared with some recent state-of-the-art HFR methods in Table 2 on the three datasets. The compared state-of-the-art HFR methods include both traditional handcrafted feature-based methods as well as deep learning-based methods. In particular, the performance of handcrafted feature-based methods, such as KDSR (Huang et al. 2013), H2(LBP3) (Shao and Fu 2017), Gabor+RBM (Yi et al. 2015), Recon.+UDP (Juefei-Xu, Pal, and Savvides 2015), Gabor+JB (Chen et al. 2012) and Gabor+HJB (Shi et al. 2017), as well as deep learning-based methods including IDNet (Reale et al. 2016), HFR-CNN (Saxena and Verbeek 2016), Hallucination (Lezama, Qiu, and Sapiro 2017), TRIVET (Liu et al. 2016), IDR (He et al. 2017), ADFL (Song et al. 2018), CDL (Wu et al. 2018b) and W-CNN (He et al. 2018) are compared in Table 2.

For the most challenging CASIA NIR-VIS 2.0 database, it can be observed from the Table 2 that DVR performs better than the other compared methods. For fair comparisons, DVR on LightCNN-9 obtains $\mathbf{99.1}\%$ on Rank-1 accuracy and $\mathbf{98.6}\%$ on VR@FAR=0.1%, which outperforms the other state-of-the-art methods on LightCNN-9, including TRIVET (Liu et al. 2016), IDR (He et al. 2017), ADFL (Song et al. 2018), CDL (Wu et al. 2018b) and W-CNN (He et al. 2018). When the backbone is changed to LightCNN-29, DVR further gains $\mathbf{0.8}\%$ on Rank-1 accuracy and $\mathbf{1.0}\%$ on VR@FAR=0.1%. The experimental results suggest that the domain discrepancy between NIR and VIS can be reduced by DVR.

For the Oulu-CASIA NIR-VIS and BUAA-VisNir databases, since the number of samples in the training set are not large enough, Table 2 demonstrates that benefiting from disentangled latent variables modeling, DVR outperforms previous state-of-the-art method such as W-CNN (He et al. 2018) by a large margin (89.7% vs 81.5% on VR@FAR=1% on Oulu-CASIA NIR-VIS as well as 97.0% vs 96.0% on VR@FAR=1% on the BUAA-VisNir database). LightCNN-29, further improves the VR@FAR=0.1% performance by $\mathbf{19.1}\%$ and $\mathbf{4.1}\%$ on the Oulu-CASIA NIR-VIS and BUAA-VisNir databases, respectively.

Fig. 2 shows the ROC curves corresponding to TRIVET (Liu et al. 2016), IDR (He et al. 2017), ADFL (Song et al. 2018), CDL (Wu et al. 2018b), W-CNN (He et al. 2018), DVR(LightCNN-9) and DVR(LightCNN-29). It can be observed that the ROC curves corresponding to the DVR method based on both LightCNN-9 and LightCNN-29 are significantly better than all the other methods. Again, this clearly shows the significance of the proposed framework for HFR. When the False Positive Rate is larger than 0.01, the True Positive Rates of all the methods are close. When the False Positive Rate tends to be small, there are large gaps between the curves of DVR and others.

Conclusion

A framework to disentangle the NIR and VIS heterogeneous face representations, called Disentangled Variational Rep-

Table 2: Comparisons with other state-of-the-art HFR methods on the CASIA NIR-VIS 2.0 database, the Oulu-CASIA NIR-VIS database and the BUAA-VisNir database.

Method	CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS			BUAA-VisNir		
	Rank-1	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%
KDSR (Huang et al. 2013)	37.5	9.3	66.9	56.1	31.9	83.0	86.8	69.5
H2(LBP3) (Shao and Fu 2017)	43.8	10.1	70.8	62.0	33.6	88.8	88.8	73.4
Gabor+RBM (Yi et al. 2015)	86.2 ± 1.0	81.3 ± 1.8	-	-	-	-	-	-
Recon.+UDP (Juefei-Xu, Pal, and Savvides 2015)	78.5 ± 1.7	85.8	-	-	-	-	-	-
Gabor+JB (Chen et al. 2012)	89.5 ± 0.8	83.2 ± 1.0	-	-	-	-	-	-
Gabor+HJB (Shi et al. 2017)	91.6 ± 0.8	89.9 ± 0.9	-	-	-	-	-	-
IDNet (Reale et al. 2016)	87.1 ± 0.9	74.5	-	-	-	-	-	-
HFR-CNN (Saxena and Verbeek 2016)	85.9 ± 0.9	78.0	-	-	-	-	-	-
Hallucination (Lezama, Qiu, and Sapiro 2017)	89.6 ± 0.9	-	-	-	-	-	-	-
TRIVET (Liu et al. 2016)	95.7 ± 0.5	91.0 ± 1.3	92.2	67.9	33.6	93.9	93.0	80.9
IDR (He et al. 2017)	97.3 ± 0.4	95.7 ± 0.7	94.3	73.4	46.2	94.3	93.4	84.7
ADFL (Song et al. 2018)	98.2 ± 0.3	97.2 ± 0.3	95.5	83.0	60.7	95.2	95.3	88.0
CDL (Wu et al. 2018b)	98.6 ± 0.2	98.3 ± 0.1	94.3	81.6	53.9	96.9	95.9	90.1
W-CNN (He et al. 2018)	98.7 ± 0.3	98.4 ± 0.4	98.0	81.5	54.6	97.4	96.0	91.9
DVR (LightCNN-9)	99.1 ± 0.2	98.6 ± 0.2	99.3	89.7	65.8	97.9	97.0	92.8
DVR (LightCNN-29)	**99.7 ± 0.1**	**99.6 ± 0.3**	**100.0**	**97.2**	**84.9**	**99.2**	**98.5**	**96.9**

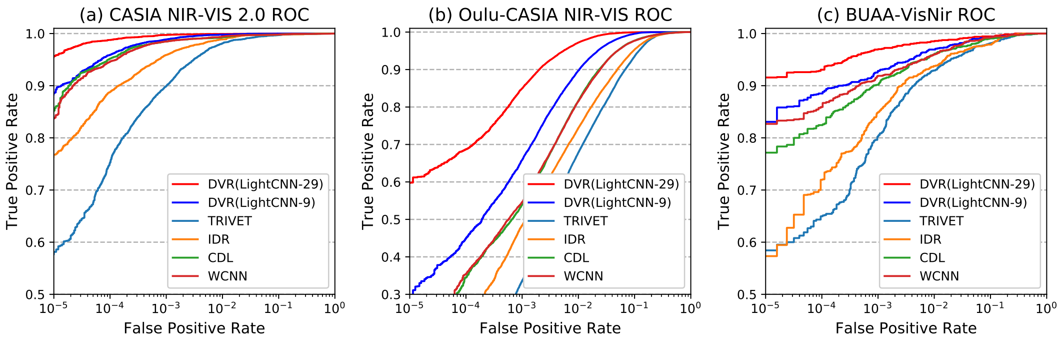

Figure 2: The ROC curves on the CASIA NIR-VIS 2.0, the Oulu-CASIA NIR-VIS and the BUAA-VisNir databases, respectively

resentation (DVR), was proposed in this paper. It provides a novel way to disentangle the NIR and VIS representations with the identity information and their within-person variations. A variational lower bound is used to estimate the posterior and optimize the disentangled latent variable space. The minimization of the identity information for the same subject and the correlation alignment constraint on the modality variations further improve the representative ability of the disentangled latent variable. An alternative optimization is employed to provide mutual promotion for both disentangled variational representation and HFR network. In this way, we can easily generate NIR and VIS samples from the likelihood according to the disentangled representations, which can effectively alleviate overfitting for HFR on the limited number of training data. Experimental results demonstrate that the proposed DVR framework leads to excellent matching accuracy on three challenging HFR databases. In addition, an ablation study is developed to demonstrate the improvements obtained by the different modules of the proposed framework.

Acknowledgments

This work is funded by National Natural Science Foundation of China (Grants No. 61622310) and Youth Innovation Promotion Association CAS (2015190).

References

Chen, J.; Yi, D.; Yang, J.; Zhao, G.; Li, S. Z.; and Pietikainen, M. 2009. Learning mappings for face synthesis from near infrared to visual light images. In *CVPR*.

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

Desjardins, G.; Courville, A. C.; and Bengio, Y. 2012. Disentangling factors of variation via generative entangling. *CoRR* abs/1210.5474.

Di, X.; Zhang, H.; and Patel, V. M. 2018. Polarimetric thermal to visible face verification via attribute preserved synthesis. In *BTAS*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial networks. In *NIPS*.

Goswami, D.; Chan, C.-H.; Windridge, D.; and Kittler, J. 2011. Evaluation of face recognition system in heterogeneous environments (visible vs nir). *ICCV Workshops*.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2017. Learning invariant deep representation for NIR-VIS face recognition. In *AAAI*.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2018. Wasserstein CNN: learning invariant features for NIR-VIS face recognition. *TPAMI.*

Hu, Y.; Wu, X.; Yu, B.; He, R.; and Sun, Z. 2018. Pose-guided photorealistic face rotation. In *CVPR*.

Huang, X.; Lei, Z.; Fan, M.; Wang, X.; and Li, S. Z. 2013. Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE TIP*.

Huang, R.; Zhang, S.; Li, T.; and He, R. 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*.

Huang, D.; Sun, J.; and Wang, Y. 2012. The BUAA-VisNir face database instructions. Technical Report IRIP-TR-12-FR-001, Beihang University, Beijing, China.

Juefei-Xu, F.; Pal, D. K.; and Savvides, M. 2015. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. *CVPR Workshops*.

Kan, M.; Shan, S.; and Chen, X. 2016. Multi-view deep network for cross-view classification. *CVPR*.

Kim, H., and Mnih, A. 2018. Disentangling by factorising. In *ICML*.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Klare, B.; Li, Z.; and Jain, A. K. 2011. Matching forensic sketches to mug shot photos. *IEEE TPAMI*.

Lei, Z.; Chu, R.; He, R.; Liao, S.; and Li, S. Z. 2007. Face recognition by discriminant analysis with gabor tensor representation. In *ICB*.

Lezama, J.; Qiu, Q.; and Sapiro, G. 2017. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. *CVPR*.

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

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

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

Liu, X.; Song, L.; Wu, X.; and Tan, T. 2016. Transferring deep representation for nir-vis heterogeneous face recognition. In *ICB*.

Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. β-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Ouyang, S.; Hospedales, T. M.; Song, Y.-Z.; and Li, X. 2016. A survey on heterogeneous face recognition: Sketch, infrared, 3d and low-resolution. *IVC*.

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

Reale, C.; Nasrabadi, N. M.; Kwon, H.; and Chellappa, R. 2016. Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition. *CVPR Workshop*.

Saxena, S., and Verbeek, J. 2016. Heterogeneous face recognition with cnns. In *ECCV Workshop*.

Schmidhuber, J. 1992. Learning factorial codes by predictability minimization. *Neural Computation*.

Shao, M., and Fu, Y. 2017. Cross-modality feature learning through generic hierarchical hyperlingual-words. *IEEE TNNLS*.

Shao, M.; Kit, D.; and Fu, Y. 2014. Generalized transfer subspace learning through low-rank constraint. *IJCV*.

Shi, H.; Wang, X.; Yi, D.; Lei, Z.; Zhu, X.; and Li, S. Z. 2017. Cross-modality face recognition via heterogeneous joint bayesian. *IEEE SPL*.

Song, L.; Zhang, M.; Wu, X.; and He, R. 2018. Adversarial discriminative heterogeneous face recognition. In *AAAI*.

Sun, B., and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*.

Tang, X., and Wang, X. 2003. Face sketch synthesis and recognition. In *ICCV*.

Wang, X., and Tang, X. 2009. Face photo-sketch synthesis and recognition. *IEEE TPAMI*.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. A. 2015. On deep multi-view representation learning. In *ICML*.

Wang, K.; He, R.; Wang, L.; Wang, W.; and Tan, T. 2016. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE TPAMI*.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018a. A light cnn for deep face representation with noisy labels. *IEEE TIFS*.

Wu, X.; Song, L.; He, R.; and Tan, T. 2018b. Coupled deep learning for heterogeneous face recognition. In *AAAI*.

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

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2015. Shared representation learning for heterogeneous face recognition. In *FG Workshops*.

Zhang, H.; Patel, V. M.; Riggan, B. S.; and Hu, S. 2017. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. In *IJCB*.

Zhao, J.; Xiong, L.; Karlekar, J.; Li, J.; Zhao, F.; Wang, Z.; Pranata, S.; Shen, S.; Yan, S.; and Feng, J. 2017. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*.

Zhu, J.-Y.; Zheng, W.-S.; Lai, J.; and Li, S. Z. 2014. Matching nir face to vis face using transduction. *IEEE TIFS*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*.

.