Shared Representation Learning for Heterogenous Face Recognition

Dong Yi, Zhen Lei, and Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences (CASIA)

***Abstract*— After intensive research, heterogenous face recognition is still a challenging problem. The main difficulties are owing to the complex relationship between heterogenous face image spaces. The heterogeneity is always tightly coupled with other variations, which makes the relationship of heterogenous face images highly nonlinear. Many excellent methods have been proposed to model the nonlinear relationship, but they apt to overfit to the training set, due to limited samples. Inspired by the unsupervised algorithms in deep learning, this paper proposes a novel framework for heterogeneous face recognition. We first extract Gabor features at some localized facial points, and then use Restricted Boltzmann Machines (RBMs) to learn a shared representation locally to remove the heterogeneity around each facial point. Finally, the shared representations of local RBMs are connected together and processed by PCA. Near infrared (NIR) to visible (VIS) face recognition problem and two databases are selected to evaluate the performance of the proposed method. On CASIA HFB database, we obtain comparable results to state-of-the-art methods. On a more difficult database, CASIA NIR-VIS 2.0, we outperform other methods significantly.**

I. Introduction

The core of heterogenous face recognition [17] is face matching across modalities. Although the original definition of heterogenous face recognition is broad, the two hottest problems about this topic are Sketch-Photo [32] and NIR-VIS (Near Infrared-Visual) [36] face recognition. Initially, heterogeneous face recognition was proposed to appeal the requirements in practical applications. Sketch-Photo matching is often required in law enforcement when the photo of suspect is unavailable. NIR-VIS matching module can make VIS face recognition system work in dark environment using NIR imaging device. After several research groups were attracted to this topic, many good methods have been proposed and these methods quickly spread to other cross-modal problems, such as face hallucination [31], pedestrian detection [33] and so on.

The Sketch-Photo problem includes two categories: view sketch recognition and forensic sketch recognition. For view sketch recognition, the sketch is drawn by an artist while viewing the subject on-site, therefore the problem is simple and very high performance has been achieved on CUFS [32] and CUFSF [38]. Unlike view sketch, forensic sketch is drawn according the description of eye-witness, thus even the best method [13] has very low performance for this problem. Compared to Sketch-Photo, the difficulty of NIR-VIS problem is bigger, thus this paper will take NIR-VIS as an example to verify the proposed method.

It has been shown in existing works that the relationship of face images between different modalities is very complex, therefore nonlinear methods usually have better performance than linear methods. Taking NIR-VIS as an example, the effect of spectrum is tightly coupled with other variations of face image, such as 3D shape, pose, identity and so on, which makes the relationship of face images under different spectrums highly nonlinear and varying with respect to locations. Among existing methods, the most successful category is learning two mappings (linear or nonlinear) to project the heterogenous face images into a common space [21][16]. Limited by the number of training samples, this kind of methods have many regularization terms, so need careful parameter tuning to achieve good performance.

From 2006 to now, unsupervised pre-training has obtained great success in deep learning [7]. One of the most popular unsupervised learning method in deep learning is Restricted Boltzmnn Machine (RBM) [26], which is a generative stochastic neural network that can learn a probability distribution of input data. To improve the generalization of existing methods and make the training process easily, this paper propose a framework based on RBM to learn the relationship of face images between different modalities. Because RBM is nonlinear and unsupervised, our framework can learn the nonlinear relationship well and unlikely prone to overfitting.

The proposed framework includes 3 main steps: (1) extracting local Gabor features around facial points, as traditional face recognition methods do; (2) learning a shared representation by RBM for each group of local features; (3) processing the whole RBM representations by PCA and matching by Cosine similarity. Among them the key step is (2), in which a 3-layer RBM is constructed and the middle layer represents the shared properties of heterogeneous data.

The contributions of this paper are as follows.

1) A local to global learning framework is proposed for heterogeneous face recognition, which can achieve good results in all experiments.
2) Local RBMs are first used to learn the shared representations of heterogenous face images. By plugging the local RBMs into the framework, we get state-of-the-art results on the CASIA HFB [19] and NIR-VIS 2.0 [18] databases.

II. Related Works

Heterogeneous face recognition research started from Tang and Wang's work in 2002 [28]. From that time to now, existing methods can be divided into two categories: Synthesis based and Classification oriented methods. In the early stage, the mainstream belongs to synthesis based methods,

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

such as [29], [22] and [30]. [29] proposed a method, named as eigen-transformation, to synthesize photo by sketch and then recognized the identity in photo modality. To get more realistic results, [22] synthesized photo in a patch way, in which each image patch was first reconstructed by LLE and then stitched into a whole photo. [30] also proposed a simple way to transform VIS to NIR face image. Although the results show that synthesis based method can achieve good visual quality, the recognition rate based on the synthesized images is moderate.

In late years, more classification oriented methods were proposed to improve the recognition rate directly. These methods just have one target: removing the difference of modalities, and meanwhile extracting discriminative feature. Many image processing and coding techniques are their essential parts, such as DoG filter [20], LBP, HOG [13], using which the difference between Sketch, NIR or VIS face images can be reduced significantly. Then, the processed heterogenous data are mapped to a discriminative space by linear, nonlinear mapping [21][16] or random trees [38]. Because the target of this kind of methods is more direct than synthesis based methods, they always perform better. In general, good methods have two properties: local processing and nonlinear mapping. The reasons behind the good methods will be analyzed in Section IV and Section V.

Recently, several methods are proposed for multi-modal problems in deep learning community. [23] first proposed a multi-modal deep learning method based on denoising autoencoder, named as Bimodal Deep AE. But the Bimodal Deep AE performs poorly in Video-Audio matching experiments. On the contrary, another shallow architecture RBM-CCA results in surprisingly good performance. Unfortunately, [23] didn't give any analysis about why the deep net was worse than RBM-CCA. In 2012, [27] pointed out that in Bimodal Deep AE the responsibility of the multi-modal modeling fell entirely on the joint layer, and other layers gave no contributions. Therefore, they proposed a multi-modal Deep Boltzmann machine (DBM), which can spread out the responsibility of the multi-modal modeling over the entire network. Experiments illustrated the superiority of DBM in Image-Text retrieval task. Then, [6] applied the multi-modal DBM in the Image-Text retrieval challenge of ICML 2013 and got the first place in the challenge.

Because the multi-modal RBM in [27] has many good properties to deal with cross-modal matching problem, we plug the multi-modal RBM into the face recognition pipeline to construct a novel method for heterogeneous face recognition. By combing the advanced modules in these two fields, the proposed method can work very well in challenging experiments.

III. Background

RBM has been widely used for modeling distribution of binary data. After Hinton's work [7], it became a standard building block of deep neural network. To model the real-valued data of face images, Gaussian RBM is used in this paper. This section will review the RBM, Gaussian RBM and Multi-modal RBM in brief.

A. Restricted Boltzmann Machines

RBM [26] is a generative stochastic graphical model that can learn the distribution of training data. The model consists of stochastic visible units $\mathbf{v} \in \{0,1\}^m$ and stochastic hidden units $\mathbf{h} \in \{0,1\}^n$, which aims to minimize the following energy function:

$$E(\mathbf{v},\mathbf{h};\mathbf{a},\mathbf{b},\mathbf{W}) = -\mathbf{a}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{v}^T\mathbf{W}\mathbf{h}, \tag{1}$$

where $\mathbf{a}$ is the biases of visible units; $\mathbf{b}$ is the biases of hidden units; $\mathbf{W}$ is the weights matrix to connect the visible and hidden units.

For image data, real-valued visible units $\mathbf{v} \in \mathbb{R}^m$ are used to replace the binary ones. The new model is called Gaussian RBM [8], the energy function of which is defined as:

$$E(\mathbf{v},\mathbf{h};\mathbf{a},\mathbf{b},\mathbf{W}) = \frac{1}{2}\mathbf{u}^T\mathbf{u} - \mathbf{b}^T\mathbf{h} - (\mathbf{v} \odot \frac{1}{\sigma})^T\mathbf{W}\mathbf{h}, \tag{2}$$

where $\mathbf{u} = (\mathbf{v}-\mathbf{a}) \odot \frac{1}{\sigma}$ denotes the normalized visible data. σ is a vector consisting of the standard deviations of each dimension. $\odot$ denotes element-wise multiplication of vectors. Before training Gaussian RBM, the input data are usually normalized by WPCA or ZCA [2], *i.e.,* the standard deviations σ of the normalized data $\hat{\mathbf{v}}$ is 1. Then, the energy function can be simplified as:

$$E(\hat{\mathbf{v}},\mathbf{h};\mathbf{a},\mathbf{b},\mathbf{W}) = \frac{1}{2}(\hat{\mathbf{v}}-\mathbf{a})^T(\hat{\mathbf{v}}-\mathbf{a}) - \mathbf{b}^T\mathbf{h} - \hat{\mathbf{v}}^T\mathbf{W}\mathbf{h}. \tag{3}$$

Then the distribution over visible and hidden units is defined as:

$$P(\hat{\mathbf{v}},\mathbf{h};\theta) = \frac{1}{Z}e^{-E(\hat{\mathbf{v}},\mathbf{h};\theta)}, \tag{4}$$

where θ is an abberation for the parameters of RBM $\{\mathbf{a},\mathbf{b},\mathbf{W}\}$; Z is a partition function defined as the sum of $e^{-E(\hat{\mathbf{v}},\mathbf{h};\theta)}$ over all possible configurations.

B. Multi-modal RBM

[27] constructed a multi-modal RBM to model the relationship between image and text by combining a Gaussian RBM and Replicated Softmax RBM. For heterogenous face recognition problem, we use two Gaussian RBM to model the relationship between face data in two modalities. The structure of our model is shown in Figure 1. Its energy function is given by:

$$\begin{aligned} E(\hat{\mathbf{v}}_1,\hat{\mathbf{v}}_2,\mathbf{h};\theta) = &\frac{1}{2}(\hat{\mathbf{v}}_1-\mathbf{a})^T(\hat{\mathbf{v}}_1-\mathbf{a}) + \\ &\frac{1}{2}(\hat{\mathbf{v}}_2-\mathbf{b})^T(\hat{\mathbf{v}}_2-\mathbf{b}) - \\ &\mathbf{c}^T\mathbf{h} - \hat{\mathbf{v}}_1^T\mathbf{W}_1\mathbf{h} - \hat{\mathbf{v}}_2^T\mathbf{W}_2\mathbf{h}, \end{aligned} \tag{5}$$

where $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$ are face images in two modalities; $\mathbf{W}_1$ and $\mathbf{W}_2$ are weights matrix for each modality respectively. The joint distribution over $\hat{\mathbf{v}}_1$, $\hat{\mathbf{v}}_2$, and $\mathbf{h}$ can be calculated based on the energy function, as similar as Eqn. (4).

Given the normalized training data $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$, we can learn the parameters θ. Then, the trained multi-modal RBM can be used flexibly, such as

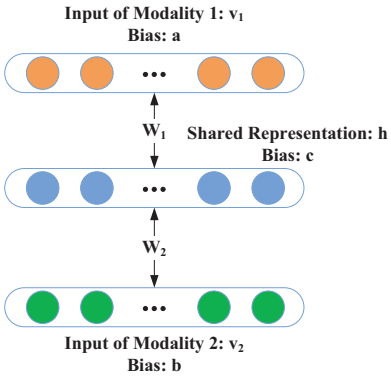

Fig. 1. A multi-modal RBM that modeling the joint distribution of face images in two modalities. The hidden layer in the model can be seen as a shared representation of the two input modalities.

1) generating missing modality by sampling from conditional distribution $P(\hat{\mathbf{v}}_1|\hat{\mathbf{v}}_2)$,
2) fusing two modalities by sampling from $P(\mathbf{h}|\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2)$,
3) inferring shared representation by sampling from $P(\mathbf{h}|\hat{\mathbf{v}}_1)$ and $P(\mathbf{h}|\hat{\mathbf{v}}_2)$ respectively.

Due to the experience in heterogeneous literature [21][16][38], this paper uses it for shared representation inference, which transforms the heterogenous data into a common space. For the details of multi-modal RBM learning and inference, please refer to [25][27].

IV. Learning Shared Representation

A. Framework

The core of heterogeneous face recognition is modeling the relationship between different modalities and meanwhile reserving the discriminative information. To this end, we propose a framework for heterogenous face recognition by incorporating RBM into the traditional face recognition pipeline. The flowchart of the framework is shown in Figure 2, in which the heterogenous face images are illustrated by NIR and VIS for example. First, Gabor features are extracted at many facial points for two modalities respectively. Based on the Gabor features, a series of local RBMs are used to learn the shared representation of two modalities for each facial point. All local shared representations are then concatenated and processed by PCA. Finally the similarity of these modality-free features can be evaluated by Cosine metric.

The proposed framework has following advantages:

1) Local Gabor feature is the mainstream in face recognition, which has strong discriminative ability and is robust to variations;
2) The shared representation are learned locally because the modality gap is smaller in local region, and low dimensional data is more efficient for computation and easier to prevent overfitting;
3) PCA can remove the redundance and heterogeneity further in holistic face.

The details of each step in Figure 2 will be described in the following subsections.

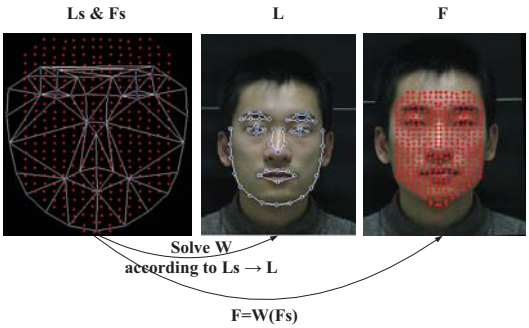

Fig. 3. The warping process of facial points. Left: Standard landmarks $\mathbf{L}_s$ (blue dots) and facial points $\mathbf{F}_s$ (red dots). Middle: A face image and its corresponding landmarks $\mathbf{L}$. Right: the warped facial points $\mathbf{F}$ for the input image.

B. Level 1 Representations

The task of level 1 is to extract discriminant and robust features for each modality. Recently, local features based on facial points achieved excellent performance in face recognition [4][35], especially in unconstrained face recognition, *e.g.,* LFW [9]. Although the face images in heterogeneous databases are both near frontal, facial points are still be used to deal with the small pose variations.

As shown in Figure 3, a standard set of facial points $\mathbf{F}_s$ are defined for feature extraction and another 48 landmarks $\mathbf{L}_s$ are defined for alignment, similar to [35]. Given a face image, we need put the facial points to the right place on it. [35] used a fast 3DMM model to do this work. For simplicity, this paper uses RBF warping [1] to transform the standard facial points to the face image. The warping process is shown in Figure 3. Given the landmarks $\mathbf{L}$ of the input image, a warping function W can be solved based on $\mathbf{L}_s$ and $\mathbf{L}$. Then the warped facial points are calculated by $\mathbf{F} = W(\mathbf{F}_s)$. We can see that the facial points can fit the input image well. The deformation factor of RBF warping is set to $0.1\times$ "eye distance".

At the warped 176×2 facial points $\mathbf{F}$, local features are extracted by a Gabor wavelet described in [24]. The space of Gabor wavelet is sampled in 8 orientations and 5 resolutions, thus giving $5 \times 8 = 40$ features for each facial point. Since the facial points are defined in a symmetric way, the features are grouped in left and right halves. Thus we get two feature vectors with 40×176 dimensions for each face image. Note that the facial symmetry trick has been used in many papers [3][19], which can augment the dataset and improve the computation efficiency.

C. Level 2 Representations

The task of level 2 is to build the relationship between two modalities. Previous work [34] has proven that the local relationship is easier to learn than holistic relationship, therefore we use local RBM to learn shared representation for each facial point. The structure of the RBMs is 40-80-40, including two input linear layers and a logistic hidden layer. Because the dimension of input of the RBM is very low, no sparse penalty and weight decay are used.

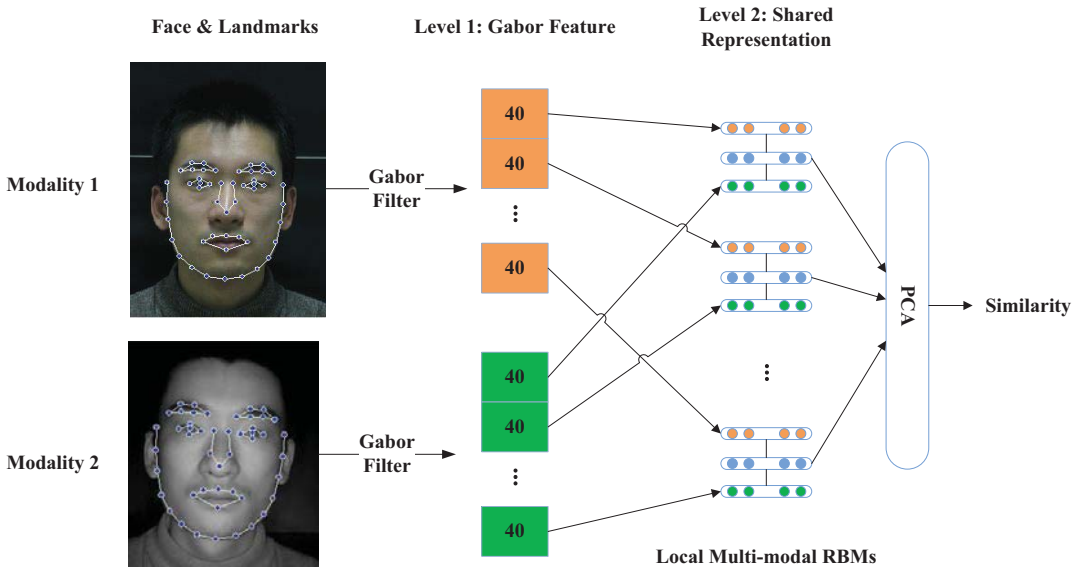

Fig. 2. The proposed framework for heterogeneous face recognition by combining traditional face recognition modules and local RBMs.

Existing methods, such as CSR [16], CITE [38] and their nonlinear versions, often learn the relationship in supervised and discriminative way. Different from them, RBM learns the joint distribution of the two modalities in a generative way, so RBM is less affected by overfitting. As described in [10], the distribution of face is not stationary with respect to the location in image, so we use many local RBMs to model the relationship between two modalities, instead of one holistic RBM.

The level 1 features of two modalities are sent to 176 local RBMs, and their parameters are learned by using mean-field inference and an MCMC procedure described in [27]. In the training stage, the batch size is set to 10 and the number of batches is set to 50000. After the training is completed, we can infer the shared representations of two modalities by sampling from $P(\mathbf{h}|\hat{\mathbf{v}}_1)$ and $P(\mathbf{h}|\hat{\mathbf{v}}_2)$. While sampling from $P(\mathbf{h}|\hat{\mathbf{v}}_1)$, we treat $\hat{\mathbf{v}}_2$ and $\mathbf{h}$ as missing data and initialize them randomly, then generate the hidden representation $\mathbf{h}$ by alternating Gibbs sampler [27]. The hidden representation of another modality can be generated in a similar way. The activation probabilities of the hidden layer are called the shared representation of heterogenous face images. The size of shared representation of a half face is 80×176.

D. Cross-modal Matching

After the heterogeneity has been removed in local regions, the heterogeneity over holistic face still exists. As described in [18], PCA can capture the heterogeneity in its first several principle components, so we use PCA to process the feature in a holistic way. First, the 176 local representations are concatenated into a vector (the dimension is $80 \times 176 = 14080$) and then the first several principle components of PCA are removed. The number of removed components is tuned on the training set or development set. To this stage, the features of two modalities are actually transformed into a common space. Their similarity can be calculated by Cosine metric. And the similarity of two face halves are fused by sum rule.

We have also tried to learn a discriminative distance metric by LDA and Metric Learning based on the shared representations, but got worse results than PCA. The reason may be due to the limited data. We believe that the supervised methods will outperform PCA when having larger database in the future.

V. Experiments

To illustrate the superior performance of the proposed method, we take NIR-VIS face recognition problem to conduct experiments. The results on two popular databases, CASIA HFB [19] and CASIA NIR-VIS 2.0 [18], all outperform the current state-of-the-art methods.

A. Databases

CASIA HFB contains 2095 VIS and 3002 NIR face images from 202 subjects. We follow the evaluation protocol in [12] that selects 102 subjects for training and the other 100 subjects for testing. The random selection is repeated in 11 times. The first split (View 1) is used to tune the parameters of algorithm, and the other 10 splits (View 2) are used to report the performance.

CASIA NIR-VIS 2.0 is an upgraded version of HFB, the images in which are captured using the same devices as HFB, but has larger scale and contains more variations in pose, facial expression and age. Compared to HFB, NIR-VIS 2.0 is more close to practical applications. This database has standard evaluation protocols, so we use them directly.

In these two experiments, VIS face images are used as gallery and NIR face images are used as probe.

B. CASIA HFB

First, Every NIR and VIS face images are processed by facial points detection and Gabor feature extraction. When

TABLE I

RANK1 RECOGNITION RATES AND VR@FAR=0.1% OF VARIOUS METHODS ON VIEW 2 OF CASIA HFB.

	Rank1	VR
Gabor	59.47±6.72%	33.51±5.70%
Gabor + Remove 20 PCs	94.87±1.72%	71.70±6.42%
Gabor + RBM	98.12±1.13%	84.50±3.75%
Gabor + RBM + Remove 11 PCs	99.38±0.32%	92.25±1.68% [1]
NN [12]	88.8%	48.78±3.87%
SR [12]	93.4%	77.56±2.96%
NN + SR [12]	92.2%	79.05±4.48%
Cognitec [12]	93.8%	85.62±2.17%
NN + SR + Cognitec [12]	97.6%	93.45±0.96%
C-DFD [15]	92.2%	65.5%
P-RS [14]	-	95.8±6.15% [2]

only using Gabor feature, we get very poor results, *i.e.,* the Rank1 recognition rate is just 50.47%. By removing the first 20 principle components of PCA, the differences between NIR and VIS are reduced significantly. The Rank1 recognition rate of Gabor+PCA increases to 94.87%, but the VR is still inferior to compared methods. The reason may be that the first 20 principle components cannot capture the difference between modalities fully. Thus we think the heterogeneity and discriminative information are coupled tightly and need to be dealt with in low level by RBM.

After introducing the RBM, the performance of our method increases significantly. As shown in Table I, the VR@FAR=0.1% is improved from 71.70% to 92.25% and the deviation is also reduced remarkably. Meanwhile, the optimal number of removed principle components drops from 20 to 11 (see Figure 4), which indicates that the modality-free representations are successfully learned by local RBMs.

Compared to other methods in [12] and [15], the Rank1 and VR of our method are obviously higher. The SR (Sparse Representation) in [12] used the whole gallery to optimize the matching process, which has been proved can improve performance, especially in terms of ROC curve. For example, the VR of our method can be improved from 92.25% to 96.33% by using z-score normalization [11]. Because in face verification applications we cannot obtain the whole gallery, we just report the results without using the whole gallery. By fusing two classifiers and a commercial face recognition SDK, the VR of NN+SR+Cognitec [12] is still lower than ours slightly. The reported performance of P-RS [14] is better than ours, but it is trained on larger training set. And P-RS is slower than our method because it's based on kernel similarities. Although CCA [37], CDFE [21] and CSR [16] are classical methods for heterogeneous face recognition, their performances are relative out-of-date (obviously lower than the numbers in Table I. The reader interested in these methods can refer Table II in [16] for details.

a) Global, Convolutional and Local RBMs: The layer in neural network has three popular styles: fully connected layer, locally connected layer with shared weights (convolutional) and locally connected layer with unshared weights (local). For RBM, we call them as global, convolutional and local RBMs. To illustrate the advantages of local RBMs, we plug them into our framework and compare their performances on View 1 of HFB, the information of which are shown in Table II. The architecture of convolutional and local RBMs are both 40-80-40. Limited by the memory of our Geforce GTX670 GPU, the hidden layer of global RBM only uses 3520 units.

[1] 96.33% with z-score normalization

[2] 133 subjects for training, 67 subjects for testing

TABLE II

THE COMPARISON OF GLOBAL, CONVOLUTIONAL AND LOCAL RBMS ON VIEW 1 OF CASIA HFB. THE 3RD COLUMN IS VR@FAR=0.1% ON THE TRAINING SET OF VIEW 1. THE 4TH COLUMN IS VR@FAR=0.1% ON THE TESTING SET OF VIEW 1.

	Architecture	VR (Train)	VR (Test)
Global	7040-3520-7040	99.94%	1.549%
Conv.	40-80-40	73.31%	71.79%
Local	176×(40-80-40)	99.45%	90.85%

The complexity of the three kinds of RBMs are global $>$ local $>$ convolutional. Generally, complex models are easier to overfit to the training set and simple models are prone to underfitting. The results in Table II prove this point well. The global RBM just performs well on the training set and the convolutional RBM performs moderately both on training and testing set. Among these models, the local RBMs obtain the best trade-off between complexity and generalization. Maybe the locality of connection and weight sharing can be fine-tuned further to get better results, but we leave this work to the future.

b) Parameter Tuning: As discussed above, the number of removed principle components greatly affects the performance of our method. Generally, if the difference between modalities is bigger, we need drop more principle components. However, there are also some identity information existing in these components, so we should find a trade-off. Figure 4 shows the relationship between the performance and the number of removed principle components on View 1. From the figure we can see that the performance of our method without RBMs are affected drastically by the first several principle components. But after using RBMs, the curves become smoother and quick to reach the optimal point, which indicates that the heterogeneities are reduced successfully in local regions by RBMs. Finally, we set the number of removed PCs to 20 when without RBMs and set the number to 11 when with RBMs.

c) Failure Cases: Although the Rank1 recognition rate of our method is very high, there are still four failure cases on View 1 of HFB, which are shown in Figure 5. From the figure we can see that the four NIR probe images both have obvious variations in pose, specular reflectance on eyeglasses and expression. Even in traditional face recognition, these factors heavily degrade the performance, thus they are more difficult to solve when coupling with spectrum variations.

C. CASIA NIR-VIS 2.0

CASIA NIR-VIS 2.0 is a more challenging and practical database than the above database. The process of this exper-

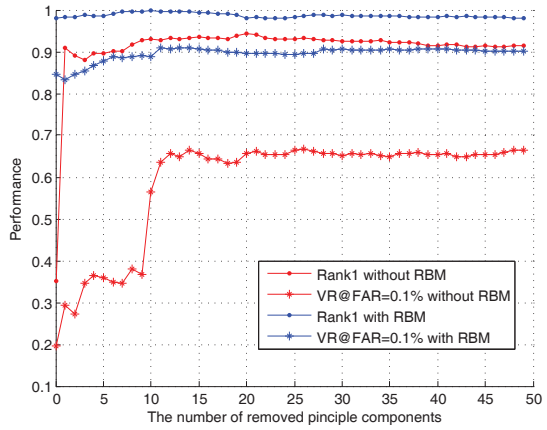

Fig. 4. The relationship between Rank1, VR and the number of removed principle components on View 1 of CASIA HFB. And the comparison curves of our method with/without RBMs.

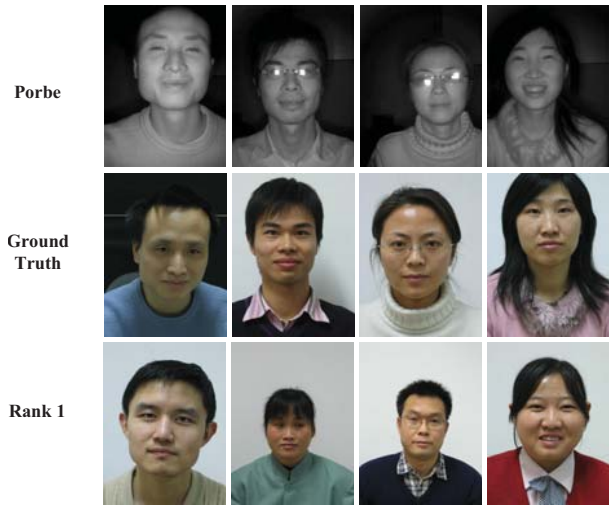

Fig. 5. The four failure cases on View 1 of CASIA HFB database. The first row are the NIR probe face images. The second row are the corresponding VIS face images of the first row. The third row are the retrieved Rank1 results of our method.

iment is as same as HFB, by first tuning the parameters on View 1 and then reporting results on View 2. From the results (Table III) we can see that the Rank1 and VR on NIR-VIS 2.0 drop 10-20% compared to HFB. But on this database, the improvements bringed by removing the first PCs and RBMs are still obvious, about 40% and 10% respectively.

Because NIR-VIS 2.0 is a relative new database, we just list the methods in [18] and [5] for comparison. The results show that the Rank1 recognition rate of the proposed method is higher than DSIFT+LDA by 12%. Because the performance of Gabor+PCA is comparable to DSIFT+LDA, we can know the improvement is mainly produced by RBMs.

Moreover, the proposed method can work well for Sketch-Photo recognition problem too. We conduct two experiments on CUFS [32] and CUFSF [38] datasets according to standard protocols. The Rank1 recognition rate is 100% on CUFS and is 98.59% on CUFSF which are on a par with the state-of-the-art results in [38].

TABLE III
RANK1 RECOGNITION RATES AND VR@FAR=0.1% OF VARIOUS METHODS ON VIEW 2 OF CASIA NIR-VIS 2.0.

	Rank1	VR
Gabor	$36.18 \pm 2.56\%$	$33.37 \pm 2.29\%$
Gabor + Remove 20 PCs	$75.54 \pm 0.75\%$	$71.40 \pm 1.21\%$
Gabor + RBM	$84.22 \pm 0.86\%$	$78.39 \pm 1.45\%$
Gabor + RBM + Remove 11 PCs	$86.16 \pm 0.98\%$	$81.29 \pm 1.82\%$
PCA + Sym + HCA [18]	$23.7 \pm 1.89\%$	19.27%
Cognitec [5]	$58.56 \pm 1.19\%$	-
DSIFT + LDA [5]	$73.28 \pm 1.10\%$	-

VI. CONCLUSION

This paper proposed a novel framework for heterogeneous face recognition by combing RBM and the popular modules from traditional face recognition methods. Because of the unsupervised nature of each module, the framework is not prone to overfitting problem, and works well on many challenging heterogeneous face databases. Based on Gabor features, the modality-free shared representations were first learned successfully in low level by many local RBMs, and further processed by PCA in high level. The proposed framework outperformed state-of-the-art methods significantly on CASIA HFB and NIR-VIS 2.0 databases. Moreover, all experimental results illustrated the success of local RBMs to learn the shared representations. The future work will be conducted in two directions: (1) by stacking many multi-modal RBMs to learn high level representations; (2) exploring the way to fine tune the model with identity information.

ACKNOWLEDGMENT

This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

REFERENCES

[1] Nur Arad and Daniel Reisfeld. "Image warping using few anchor points and radial functions". *Computer Graphics Forum*, 14(1):35–46, 1995.

[2] Anthony Bell and Terrence J. Sejnowski. "The independent components of natural scenes are edge filters". *Vision Research*, 37:3327–3338, 1997.

[3] Thomas Berg and Peter Belhumeur. "Tom-vs-Pete classifiers and identity-preserving alignment for face verification". In *Proceedings of the British Machine Vision Conference*, pages 129.1–129.11, 2012.

[4] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification". In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032, 2013.

[5] 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 *Proceedings of International Conference on Pattern Recognition*, Stockholm, Sweden, August 2014.

[6] Fangxiang Feng, Ruifan Li, and Xiaojie Wang. "Constructing hierarchical image-tags bimodal representations for word tags alternative choice". *CoRR*, abs/1307.1275, 2013.

[7] G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". *Science*, 313(5786):504–507, 28 July 2006.

[8] Geoffrey E. Hinton. "A practical guide to training restricted boltzmann machines". In Grgoire Montavon, GeneviveB. Orr, and Klaus-Robert Mller, editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 599–619. Springer Berlin Heidelberg, 2012.

[9] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. "Labeled faces in the wild: A database for studying face recognition in unconstrained environments". Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[10] G.B. Huang, Honglak Lee, and E. Learned-Miller. "Learning hierarchical representations for face verification with convolutional deep belief networks". In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525, 2012.

[11] Anil Jain, Karthik Nandakumar, and Arun Ross. "Score normalization in multimodal biometric systems". *Pattern Recognition*, 38:2270–2285, 2005.

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

[13] B.F. Klare, Zhifeng Li, and A.K. Jain. "Matching forensic sketches to mug shot photos". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(3):639–646, 2011.

[14] Brendan F Klare and Anil K Jain. Heterogeneous face recognition using kernel prototype similarities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(6):1410–1422, 2013.

[15] Z. Lei, M. Pietikainen, and S. Li. "Learning discriminant face descriptor". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99), 2013.

[16] Zhen Lei, Shengcai Liao, A.K. Jain, and S.Z. Li. "Coupled discriminant analysis for heterogeneous face recognition". *IEEE Transactions on Information Forensics and Security*, 7(6):1707–1716, 2012.

[17] Stan Z. Li and Anil K. Jain, editors. *The Encyclopedia of Biometrics*. 2009.

[18] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. "The CASIA NIR-VIS 2.0 face database". In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013.

[19] S.Z. Li, Zhen Lei, and Meng Ao. "The HFB face database for heterogeneous face biometrics research". In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2009.

[20] Shengcai Liao, Dong Yi, Zhen Lei, Rui Qin, and StanZ. Li. "Heterogeneous face recognition from local structures of normalized appearance". In Massimo Tistarelli and MarkS. Nixon, editors, *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 209–218. 2009.

[21] Dahua Lin and Xiaoou Tang. "Inter-modality face recognition ". In *Proceedings of the European Conference on Computer Vision*, volume 3954, pages 13–26, 2006.

[22] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. "A nonlinear approach for face sketch synthesis and recognition". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1005–1010, 2005.

[23] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. "Multimodal deep learning". In *ICML*, pages 689–696, 2011.

[24] Kazunori Okada, Johannes Steffens, Thomas Maurer, Hai Hong, Egor Elagin, Hartmut Neven, and Christoph von der Malsburg. "The Bochum/USC Face Recognition System and How it Fared in the FERET Phase III Test", 1998.

[25] Ruslan Salakhutdinov and Geoffrey E. Hinton. "Deep boltzmann machines". *Journal of Machine Learning Research - Proceedings Track*, 5:448–455, 2009.

[26] P. Smolensky. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. chapter Information processing in dynamical systems: foundations of harmony theory, pages 194–281. MIT Press, Cambridge, MA, USA, 1986.

[27] Nitish Srivastava and Ruslan Salakhutdinov. "Multimodal learning with deep boltzmann machines". In *Proceedings of Neural Information Processing Systems*, pages 2231–2239, 2012.

[28] Xiaoou Tang and Xiaogang Wang. "Face photo recognition using sketch". In *International Conference on Image Processing*, volume 1, pages 257–260, 2002.

[29] Xiaoou Tang and Xiaogang Wang. "Face sketch synthesis and recognition". In *IEEE International Conference on Computer Vision*, volume 1, pages 687–694, 2003.

[30] Rui Wang, Jimei Yang, Dong Yi, and StanZ. Li. "An analysis-by-synthesis method for heterogeneous face biometrics". In *Proceedings of IAPR International Conference on Biometric*, pages 319–326, 2009.

[31] Xiaogang Wang and Xiaoou Tang. "Hallucinating face by eigentransformation". *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 35(3):425–434, 2005.

[32] Xiaogang Wang and Xiaoou Tang. "Face photo-sketch synthesis and recognition". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):1955–1967, 2009.

[33] Junjie Yan, Xucong Zhang, Zhen Lei, Dong Yi, Shengcai Liao, and Stan Z. Li. "Robust multi-resolution pedestrian detection in traffic scenes". In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[34] Weilong Yang, Dong Yi, Zhen Lei, Jitao Sang, and S.Z. Li. "2D-3D face matching using cca". In *IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, 2008.

[35] Dong Yi, Zhen Lei, and Stan Z. Li. "Towards pose robust face recognition". In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3539–3545, 2013.

[36] Dong Yi, Rong Liu, Rufeng Chu, Zhen Lei, and Stan Z. Li. "Face matching between near infrared and visible light images". In *Proceedings of IAPR International Conference on Biometric*, Seoul, Korea, August 2007.

[37] Dong Yi, Rong Liu, Rufeng Chu, Zhen Lei, and Stan Z. Li. Face matching between near infrared and visible light images. In *ICB*, pages 523–530, 2007.

[38] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. "Coupled information-theoretic encoding for face photo-sketch recognition". In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 513–520, 2011.