The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Asymmetric Joint Learning for Heterogeneous Face Recognition

Bing Cao,[1] Nannan Wang,[2*] Xinbo Gao[1] Jie Li[1]
[1]State Key Laboratory of Integrated Services Networks,
School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]State Key Laboratory of Integrated Services Networks,
School of Telecommunications, Xidian University, Xi'an 710071, China

Abstract

Heterogeneous face recognition (HFR) refers to matching a probe face image taken from one modality to face images acquired from another modality. It plays an important role in security scenarios. However, HFR is still a challenging problem due to great discrepancies between cross-modality images. This paper proposes an asymmetric joint learning (AJL) approach to handle this issue. The proposed method transforms the cross-modality differences mutually by incorporating the synthesized images into the learning process which provides more discriminative information. Although the aggregated data would augment the scale of intra-classes, it also reduces the diversity (*i.e.* discriminative information) for inter-classes. Then, we develop the AJL model to balance this dilemma. Finally, we could obtain the similarity score between two heterogeneous face images through the log-likelihood ratio. Extensive experiments on viewed sketch database, forensic sketch database and near infrared image database illustrate that the proposed AJL-HFR method achieve superior performance in comparison to state-of-the-art methods.

1 Introduction

Heterogeneous face images refer to images that represent faces in different modalities, such as sketch images (drawn by artists), visual (VIS) images (captured through general camera) and near infrared (NIR) images (captured through near infrared devices). Matching face images between different modalities is called heterogeneous face recognition (HFR), which is an important issue in security scenarios. For instance, the photos of suspects are usually difficult to obtain during the law enforcement process. Then the HFR is desired to identify suspects by matching sketches drawn by artists with photos in mug-shot databases. And HFR is also aimed to matching NIR images with VIS images when the circumambient illumination condition is poor for face recognition in public security.

Due to the great differences between heterogeneous face images, it is difficult for conventional face recognition methods to identify a face sketch or a near infrared face image from visual face photos. Existing HFR methods can be grouped into three categories: local feature descriptor-based methods, subspace learning-based methods and synthesis-based methods.

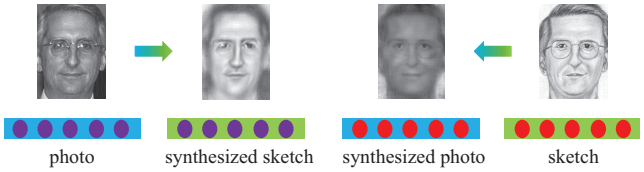

Figure 1: The latent information in synthesized images. Different colors of circles and backgrounds represent different texture information and domains respectively. For example, The purple circles with the blue rectangle background represent the texture information of one subject in photo domain and the purple circles with the green rectangle background represent the same texture information of the synthesized sketch in sketch domain.

Local feature descriptor-based methods extract local feature descriptors to represent heterogeneous face images, which is desired to reduce the discrepancies between heterogeneous face images in feature level. Yet, because of the high computational complexity and limited discriminability, these encoded feature descriptors take much time and perform poor in recognition tasks. Subspace learning-based methods project face images from different modalities into a common subspace to minimize the discrepancies. Within this subspace, heterogeneous face images can be matched directly. But, it is inevitable to lose effective information in the projection procedure, which decreases recognition performance. Synthesis-based methods train a set of reconstruction coefficients to transform heterogeneous face images to homogeneous face images, which is aimed to reduce the discrepancies between heterogeneous face images in image level. Then these homogeneous face images can be directly applied to conventional face recognition approaches. However, the synthesized photos only change the color of corresponding sketches, which could never reduce the differences in texture (*e.g.* the shape of glasses frame, double-fold eyelid and facial outline) between heterogeneous face images and achieve poor performance in HFR. Recently, an efficient joint formulation (Chen et al. 2017) for conventional

*Corresponding author: Nannan Wang (nnwang@xidian.edu.cn)

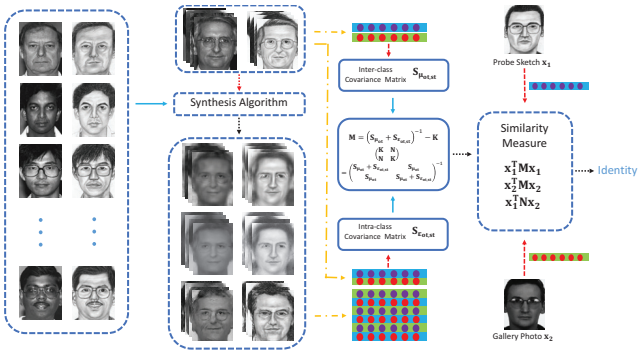

Figure 2: Framework of the proposed asymmetric joint learning method for heterogeneous face recognition.

face recognition was proposed, which enhanced the original Bayesian face model and achieved a promising result. This method takes intra-personal variations and inter-personal invariants over the image pairs into consideration and requires that there are enough intra-class face images. But most of heterogeneous face databases provide few intra-class face images and the differences between heterogeneous face images are cross-modality. So the performance of this method on heterogeneous face databses is not good.

This paper proposes a novel asymmetric joint learning approach for heterogeneous face recognition (AJL-HFR). The proposed method takes the latent information from synthesized face images into consideration. The latent information refers to texture of synthesized face images. Note that, the synthesized images and corresponding original face images together provide more information of the same texture in different modalities, which is shown in Figure. 1. When extracting intra-class information, we add the synthesized face images into the training set. Therefore, we could acquire more effective intra-class information from the enlarged training set. However, it also introduces redundant variables to inter-class information, which influences the recognition performance. To balance this dilemma, we design an asymmetric joint learning model to extract more effective intra-class information without losing inter-class information in the training phase. Two covariance matrices that represent two kinds of information are jointly optimized by original images and synthesized images. Finally, the log-likelihood ratio statistic is calculated as similarity score of two input heterogeneous face images. The outline can be found from Figure. 2. We evaluate our methods on four databases: CUHK Face Sketch FERET (CUFSF) database (Zhang, Wang, and Tang 2011), IIIT-D Sketch database (Bhatt et al. 2012), Forensic Sketch database (Peng et al. 2017) and CUHK VIS-NIR database (Gong et al. 2017). Figure. 3 shows some samples in these databases.

The contributions of this paper are summarized as follows:

- We firstly utilize the latent information involved in synthesized images to extract more effective intra-class information and design a valid strategy to obtain more information from the limited databases.

Figure 3: The illustration of heterogeneous face databases. (a): CUFSF database. (b):IIIT-D Sketch database. (c): Forensic Sketch database. (d): CUHK VIS-NIR database.

- An asymmetric joint learning model is developed to jointly optimize intra-class and inter-class information without losing effective inter-class information.
- Experimental results illustrate the superior performance of our proposed approach compared with state-of-the-art HFR methods on multiple HFR scenarios.

We organize the rest of this paper as follows. Section 2, representative heterogeneous face recognition methods are briefly reviewed. Section 3, the proposed asymmetric joint learning approach is introduced in detail. We provide some experimental results and analysis in section 4, and summarize this paper in section 5.

2 Related Work

In this section, we briefly review some representative HFR methods in three categories: local feature descriptor-based methods, subspace learning-based methods and synthesis-based methods.

Local feature descriptor-based methods focus on extracting the invariant features from face images in different modalities. Most of these methods utilized local features for HFR, such as LFDA (Klare, Li, and Jain 2011), CITE (Zhang, Wang, and Tang 2011), LRBP (Galoogahi and Sim 2012), MCWLD (Bhatt et al. 2012), LDoGBP (Alex, Asari, and Mathew 2013), and CEFD (Gong et al. 2017). Besides, Mittal, Vatsa, and Singh (2015) and He et al. (2017) employed deep learning to extract invariant features for HFR. To summarize, feature-based methods try to minimize modality discrepancies in heterogeneous face images. However, due to the high computational complexity and limited discriminability, the accuracies of these methods still remain to improve.

Subspace learning-based methods aim to project heterogeneous facial features into a common subspace. A general discriminant feature extraction method was firstly proposed in Lin and Tang (2006), which can transform the features in different modalities into the common space respectively. Sharma and Jacobs (2011) proposed partial least squares (PLS) approach to get a common linear subspace. Lei et al. (2012) applied coupled spectral regression method in the projection procedure. A non-linear kernel was applied to represent heterogeneous face images in P-RS (Klare and Jain 2013). Kan et al. (2016) utilized the relationship of cross-modality facial images to develop a multi-view discriminant analysis (MvDA) approach. Actually, it is unavoidable to lose some discriminative information in the pro-

jection procedure, and leads to dissatisfactory performance in HFR.

Synthesis-based methods try to synthesize heterogeneous face images and compare them in the same modality. These methods are almost based on a set of reconstruction coefficients and heterogeneous image patches (Chen et al. 2009; Wang and Tang 2009; Zhou, Kuang, and Wong 2012; Gao et al. 2012; Zhang et al. 2011; Wang et al. 2013; Song et al. 2014; Wang, Gao, and Li 2018). In addition, we introduce generative adversarial networks (GANs) (Isola et al. 2016) to synthesize heterogeneous face images. However, each method has its own pro and con. Though synthesis-based methods can reduce the modality discrepancies in heterogeneous face images, the synthesis procedure takes a long time, which slows down the HFR process. Meanwhile, images synthesis is another difficult issue. Therefore, synthesis-based methods perform unpleasantly for HFR. *However, the synthesized face sketches reflect different aspects of photos, which is the basic motivation of our proposed approach. Synthesized sketches and photos are only used for training in our proposed method which thus does not increase the online time-consuming for HFR.*

3 Asymmetric Joint Learning

We present a novel framework for HFR in this section, which is called asymmetric joint learning for heterogeneous face recognition (AJL-HFR). Without loss of generality and for ease of representation, we describe our approach on face sketch-photo recognition scenario, which can be generalized to other heterogeneous face recognition scenarios. Firstly, we introduce our motivation. Then we introduce how to derive the proposed model and how to optimize it.

3.1 Motivation

We find that the texture information (*e.g.* the shape of glasses frame, double-fold eyelid and facial outline) of the same subject in the photo domain and the sketch domain is different. For example, we list some photo-sketch pairs in Figure. 4. Among them, the first column is the original photo-sketch pair, followed by synthesized photo-sketch pairs generated by Chen et al. (LLE), Wang and Tang (MRF), Zhou, Kuang, and Wong (MWF), Gao et al. (SFS), Zhang et al. (SVR), Gao et al. (SRE), Wang et al. (RSLCR), and GANs. In the original photo-sketch pair, the glasses in the photo are frameless, but the artists draw the glasses frame in the sketch to incarnate the glasses clearly. There are many similar differences like this between heterogeneous images, such as hair, wrinkles and so on. Although these differences seriously affect the final recognition performance, we find some interesting phenomenon from these differences.

Since we utilize the linear combination of the reconstructed coefficients of the sketches and the photo patches in the training set to synthesize photos (Wang et al. 2014), the texture of synthesized photos are more similar to the original sketches than original photos. Similarly, the texture of synthesized sketches are more similar to the original photos than original sketches. Thus, through photos and the corresponding synthesized sketches, we can obtain more information

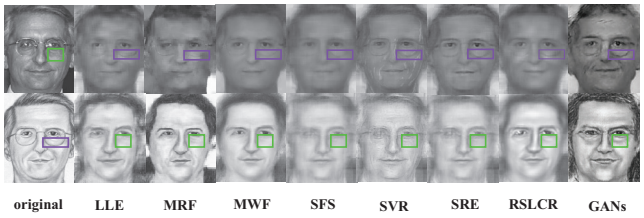

Figure 4: Original photo-sketch pairs and the corresponding synthesized photo-sketch pairs. The names of the synthesis methods are presented under the image pairs.

of same texture in different modalities. As shown in Figure. 4, if these synthesis methods are conditionally selected to enlarge the training set, we can obtain more useful information. However, the selected synthesis images should not be too many for these images contain both valid information and redundant information. When there are too many synthesized images, the redundant information could overweigh the valid information. In addition, the inter-class information is obtained from the intra-class averages of all classes. If synthesized images are added into the training set, the effectiveness of inter-class information could be reduced. To balance this dilemma of valid information and redundant information, we propose the asymmetric joint learning framework for heterogeneous face recognition.

3.2 Model Derivation

As shown in Figure. 2, we need generate some pairs of synthesized sketch and synthesized photo to construct the training database for learning the AJL model. Here we generated three synthesized sketch-photo pairs by RSLCR, MWF and GANs respectively. These three methods are chosen from three different categories of face sketch synthesis methods respectively. The rationale behind this selection strategy would be given in section 4. We extract CNN features from synthesized image pairs and original image pair to represent each subject. AJL model (the notation $\mathbf{M}$ and $\mathbf{N}$ in Figure. 2) is calculated from the inter-class covariance matrix and intra-class covariance matrix. In the recognition phase, the similarity between the input query image and the gallery image is calculated by a log-likelihood ratio based on the trained AJL model.

Inspired by the metric learning model (Chen et al. 2017), a face $\mathbf{x}$ can be approximated by inter-class variation μ and intra-class variation ε, where μ represents the identity and ε represents the heterogeneous information between cross-modality face images belonging to the same identity. Similar with the preceding models (Belhumeur, Hespanha, and Kriegman 1997; Ioffe 2006; Susskind et al. 2011), these two parts are modeled by independent zero-mean Gaussian,

$$\begin{aligned} \mu &\sim \mathcal{N}(0, \mathbf{S}_\mu), \\ \varepsilon &\sim \mathcal{N}(0, \mathbf{S}_\varepsilon), \end{aligned} \tag{1}$$

where $\mathbf{S}_\mu$ and $\mathbf{S}_\varepsilon$ are two covariance matrices to be trained. A face image can be represented by the two parts as follows,

$$\mathbf{x} = \mu + \varepsilon. \tag{2}$$

For the two input face images $\mathbf{x}_1$ and $\mathbf{x}_2$, the covariance of which can be written as

$$\mathbf{cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{cov}(\mu_1, \mu_2) + \mathbf{cov}(\varepsilon_1, \varepsilon_2). \quad (3)$$

If $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to the same identity, denoted by $\mathbf{H_I}$, or else, donated by $\mathbf{H_E}$. Thus, the intra-class joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_I})$ is a Gaussian with zero-mean and the covariance matrix of which is

$$\sum_{\mathbf{I}} = \begin{bmatrix} \mathbf{S}_\mu + \mathbf{S}_\varepsilon & \mathbf{S}_\mu \\ \mathbf{S}_\mu & \mathbf{S}_\mu + \mathbf{S}_\varepsilon \end{bmatrix}. \quad (4)$$

The inter-class joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_E})$ is a Gaussian with zero-mean and the covariance matrix of which is

$$\sum_{\mathbf{E}} = \begin{bmatrix} \mathbf{S}_\mu + \mathbf{S}_\varepsilon & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\mu + \mathbf{S}_\varepsilon \end{bmatrix}. \quad (5)$$

The covariance matrix $\mathbf{S}_\mu$ is trained from the combination of the average of each class. And the covariance matrix $\mathbf{S}_\varepsilon$ is trained from all the samples of each class. However, for heterogeneous face images, the images belonging to the same subject come from different modalities. Therefore, it is not sufficient to extract intra-class information by only modeling the single pair of sketch-photo images in the training dataset which is adopted by joint Bayesian method (Chen et al. 2017). In order to obtain more useful intra-class information of heterogeneous face images, we jointly train the model relying on both the original training dataset and the synthesized image pairs. Thus, the covariance matrix $\mathbf{S}_\varepsilon$ provides more information about intra-class face images. However, since the synthesized image pairs are pseudo heterogeneous face image pairs, they are different from the image pairs in the training set and would bring in redundant information. When we add all these synthesized image pairs (*e.g.* Figure. 4) to the training set and train a joint Bayesian model directly, the covariance matrices $\mathbf{S}_\mu$ and $\mathbf{S}_\varepsilon$ would be contaminated by these redundant information inevitably. It seriously affects the effectiveness of joint Bayesian model. To solve this problem and extract more useful intra-class and inter-class information, we improve the model by an asymmetric joint learning model.

The asymmetric joint learning model firstly generates a certain number of sketch-photo pairs. We present the detailed illustration about how to generate synthesized sketches and photos in section 4. Then, the intra-class and inter-class covariance matrices are trained jointly. The intra-class covariance matrix $\mathbf{S}_{\varepsilon_{\mathbf{ot,st}}}$ is derived from the original training set (*denoted as* $\mathbf{ot}$) and the corresponding synthesized images (*denoted as* $\mathbf{st}$). The inter-class covariance matrix $\mathbf{S}_{\mu_{\mathbf{ot}}}$ is derived only from original training set. For the two covariance matrices are independent, the covariance matrix of intra-class joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_I})$ can be written as

$$\sum_{\mathbf{I}} = \begin{bmatrix} \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & \mathbf{S}_{\mu_{\mathbf{ot}}} \\ \mathbf{S}_{\mu_{\mathbf{ot}}} & \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} \end{bmatrix}. \quad (6)$$

The covariance matrix of the inter-class joint distribution $\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_E})$ can be written as

$$\sum_{\mathbf{E}} = \begin{bmatrix} \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} \end{bmatrix}. \quad (7)$$

Finally, omitting the constant parameter, we can compute the log-likelihood ratio $r(\mathbf{x}_1, \mathbf{x}_2)$ to obtain the similarity of two input cross-modality face images by intra-class joint distribution and inter-class joint distribution as

$$\begin{aligned} r(\mathbf{x}_1, \mathbf{x}_2) &= \log \frac{\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_I})}{\mathbf{P}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{H_E})} \\ &= \mathbf{x}_1^T\mathbf{M}\mathbf{x}_1 + \mathbf{x}_2^T\mathbf{M}\mathbf{x}_2 - 2\mathbf{x}_1^T\mathbf{N}\mathbf{x}_2, \end{aligned} \quad (8)$$

where

$$\mathbf{M} = (\mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}})^{-1} - \mathbf{K}, \quad (9)$$

and $\mathbf{K}$ satisfies

$$\begin{bmatrix} \mathbf{K} & \mathbf{N} \\ \mathbf{N} & \mathbf{K} \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & \mathbf{S}_{\mu_{\mathbf{ot}}} \\ \mathbf{S}_{\mu_{\mathbf{ot}}} & \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} \end{bmatrix}^{-1}. \quad (10)$$

Therefore, we turn the face verification issue into estimating the two covariance matrices $\mathbf{S}_{\mu_{\mathbf{ot}}}$ and $\mathbf{S}_{\mu_{\mathbf{ot,st}}}$.

Suppose that there are m_i $i.i.d.$ intra-class face images belonging to the same subject i. According to equation (2), we can represent all the samples of the same subject by

$$\mathbf{x}_i = \mathbf{Q}_i\mathbf{h}_i, \quad (11)$$

where

$$\begin{aligned} \mathbf{Q}_i &= \begin{bmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}, \\ \mathbf{h}_i &= \left[\mu_{i_{ot}}; \varepsilon_{i_{ot,st}1}; \varepsilon_{i_{ot,st}2}; \cdots; \varepsilon_{i_{ot,st}m_i}\right], \end{aligned} \quad (12)$$

and $\mathbf{I}$ is the identity matrix whose dimension is determined by the feature dimension.

Note that, our objective function is

$$\max \textstyle\prod_i \mathbf{P}(\mathbf{x}_i|\mathbf{h}_i). \quad (13)$$

For each identity, the inter-class variation $\mu_{i_{ot}}$ can be derived from $\mathcal{N}(0, \mathbf{S}_{\mu_{\mathbf{ot}}})$. Then the intra-class variations $\left[\varepsilon_{i_{ot,st}1}; \varepsilon_{i_{ot,st}2}; \cdots; \varepsilon_{i_{ot,st}m_i}\right]$ can be derived from $\mathcal{N}(0, \mathbf{S}_{\mu_{\mathbf{ot,st}}})$. Because subjects are independent, the objective function is equivalent to

$$\max \textstyle\sum_i \log \mathbf{P}(\mathbf{x}_i|\mathbf{S}_{\mu_{\mathbf{ot}}}, \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}}). \quad (14)$$

To solve this problem, we develop an asymmetric EM algorithm to jointly optimize the estimation of two covariance matrices in next subsection.

3.3 Model Optimization

E-Step Considering the distribution of the latent variation $\mathbf{h}_i$ is a Gaussian with the covariance matrix,

$$\sum_{\mathbf{h}_i} = \begin{bmatrix} \mathbf{S}_{\mu_{\mathbf{ot}}} & & & \\ & \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & & \\ & & \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & \\ & & & \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} \end{bmatrix}, \quad (15)$$

we can write the likelihood function of subject i as

$$\mathbf{P}(\mathbf{x}_i|\mathbf{S}_{\mu_{\mathbf{ot}}}, \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}}) = \mathcal{N}(\mathbf{0}, \textstyle\sum_{\mathbf{x_i}}), \quad (16)$$

Algorithm 1 AJL-HFR

Input: Training set $\mathbf{A}$, probe image $\mathbf{p}$, gallery dataset $\mathbf{G}$.

***Step* 1**: Generate synthesized image pairs corresponding to training set $\mathbf{A}$ by three face sketch synthesis methods: RSLCR, MWF, GANs. Let $\mathbf{B}$ represents the set of the synthesized image pairs and the original training image pairs.

***Step* 2**: Initialize the inter-class covariance matrix $\mathbf{S}_{\mu_{\mathbf{ot}}}$ from image pairs of training set $\mathbf{A}$ and the intra-class covariance matrix $\mathbf{S}_{\varepsilon_{\mathbf{ot,st}}}$ from image pairs of dataset $\mathbf{B}$.

***Step* 3**: EM strategy is applied to jointly optimize $\mathbf{S}_{\mu_{\mathbf{ot}}}$ and $\mathbf{S}_{\varepsilon_{\mathbf{ot,st}}}$. Then calculate $\mathbf{M}$ and $\mathbf{N}$ according to equation (9) and (10) respectively.

***Step* 4**: Calculate the similarity of probe image $\mathbf{p}$ and each image in gallery dataset $\mathbf{G}$. Sort the similarities by descend order.

Output: The target heterogeneous face image $\mathbf{t}$ in gallery dataset $\mathbf{G}$.

where

$$\sum_{\mathbf{x}_i} = \mathbf{Q}_i \sum_{\mathbf{h}_i} \mathbf{Q}_i^T = \begin{bmatrix} \mathbf{I} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{S}_{\mu_{\mathbf{ot}}} & & & \\ & \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & & \\ & & \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & \\ & & & \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{I} & \cdots & \mathbf{I} \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & \mathbf{S}_{\mu_{\mathbf{ot}}} & \cdots & \mathbf{S}_{\mu_{\mathbf{ot}}} \\ \mathbf{S}_{\mu_{\mathbf{ot}}} & \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} & \cdots & \mathbf{S}_{\mu_{\mathbf{ot}}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{\mu_{\mathbf{ot}}} & \mathbf{S}_{\mu_{\mathbf{ot}}} & \cdots & \mathbf{S}_{\mu_{\mathbf{ot}}} + \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} \end{bmatrix} \tag{17}$$

Based on the learning process in Chen et al. (2017), the expectation of latent variable $\mathbf{h}_i$ can be computed by

$$\mathbf{E}_{\mathbf{P}(\mathbf{h}_i | x_i, \mathbf{S}_{\mu_{\mathbf{ot}}}, \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}})} = \sum_{\mathbf{h}_i} \mathbf{Q}_{\mathbf{i}}^{\mathbf{T}} \sum_{\mathbf{x}_i}^{-1} \mathbf{x}_i. \tag{18}$$

At the beginning of E-Step, we asymmetrically initialize $\mathbf{S}_{\mu_{\mathbf{ot}}}$ by the covariance of the mean of each inter-class identity from original training set, and initialize $\mathbf{S}_{\varepsilon_{\mathbf{ot,st}}}$ by the covariance of intra-class face images from original training set and synthesized face images.

M-Step As the latent variable $\mathbf{h}_i$ has been estimated in the last step, we can update the parameters by substituting $\mu_{i_{ot}}$ and $\varepsilon_{i_{ot}}$ into the following equation:

$$\begin{aligned} \mathbf{S}_{\mu_{\mathbf{ot}}} &= \tfrac{1}{n} \textstyle\sum_i \mathbf{E}\left[\mu_{i_{ot}} \mu_{i_{ot}}^T\right], \\ \mathbf{S}_{\varepsilon_{\mathbf{ot,st}}} &= \frac{\sum_i \sum_j \mathbf{E}\left[\varepsilon_{i_{ot,st}j} \varepsilon_{i_{ot,st}j}^T\right]}{\sum_i m_i}, \end{aligned} \tag{19}$$

where n represents the number of subjects in training set.

The algorithm generally converges in fifty iterations, then we can utilize the equations (8)-(10) to compute the similarities between the probe image and the gallery images. Algorithm 1 summarizes the implementation steps for the proposed asymmetric joint learning method for heterogeneous face recognition.

4 Experimental Results and Analysis

In this section, we validated the effectiveness of the proposed approach on four HFR scenarios, *i.e.* viewed sketches vs. visible images, semi-forensic sketches vs. visible images, forensic sketches vs. visible images and near infrared images vs. visible images. To begin with, we explored the effect of different features and different combinations of face sketch synthesis methods. Then, we confirmed the superior performance of our approach compared with state-of-the-art methods on CUFSF database, IIIT-D Sketch database, Forensic Sketch database and CUHK VIS-NIR database.

4.1 Databases and Protocols

We present four HFR scenarios as shown in Figure. 3. The viewed sketches are drawn by artists as viewing photos. For viewed sketch database, we use the CUFSF database which contains 1194 sketch-photo pairs and 500 subjects are randomly selected as the training set. The remaining 694 subjects are used for test.

The forensic sketches are generally used for law enforcement which are drawn according to descriptions of eyewitnesses or victims. There are great gap between the viewed sketches and the forensic sketches. To decrease the gap, researchers develop semi-forensic sketches (IIIT-D Sketch database) which are drawn by artists according to their memory to the photos which are viewed once to the artists. There are two partition protocols for these databases. One protocol is training on semi-forensic database, and then testing on the forensic sketch database. The forensic sketch database contains 168 mug shot photos and corresponding forensic sketches from real world. We follow the same protocol in Peng et al. (2017), *i.e.* 124 sketch-photo pairs of IIIT-D database are randomly selected as the training set. Face sketch synthesis models are trained on CUHK AR sketch database (Wang and Tang 2009). Then we evaluate the performance on 168 real forensic sketch-photo pairs. The other protocol is training and testing both on the forensic sketch database. We randomly select 112 subjects as the training set. 250 subjects from CUFSF database are randomly selected to generate synthesized image pairs. The remaining 56 subjects are used as the testing set. The gallery sets are all extended by 10000 photos that are randomly selected from LFW database (Huang et al. 2007).

The near infrared images are formed by the reflected infrared waves of objects. For this HFR scenario, we evaluate the proposed approach on CUHK VIS-NIR database (Gong et al. 2017), each subject in which has one pair of near infrared image-visible image. There are 2800 subjects in total. Considering the partition protocol in Gong et al. (2017), we randomly divide the database into two halves without overlapping, one half for training and the other half for testing.

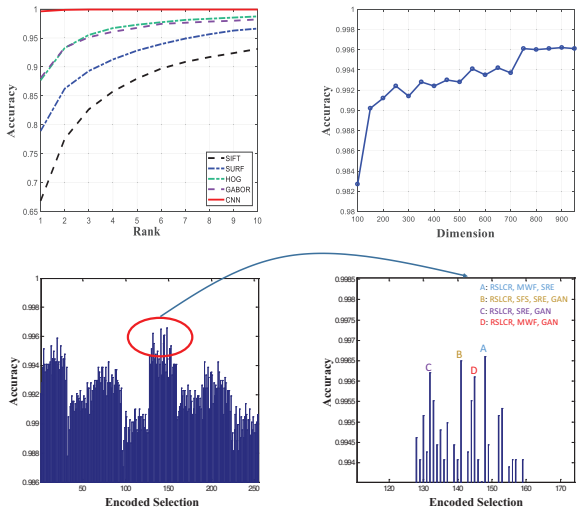

Figure 5: Left top subfigure gives the performance comparison with different features; right top subfigure illustrates the influence with different feature dimensions; left bottom subfigure provides the results corresponding to different combinations of eight synthesis methods (there are totally 2^8 different combinations). The horizental axis is the decimal digit corresponding to the binary number, and the vertical axis is the rank-1 recognition accuracy corresponding to different combinations. There we set 'RSLCR', 'LLE', 'MRF', 'MWF', 'SFS', 'SRE', 'SVR' and 'GANs' as a binary digit with 8 bits. Each method is corresponding to one bit. For instance, the binary digit '10101010' represent the combination of 'RSLCR', 'MRF', 'SFS' and 'SVR'. And the corresponding decimal digit is 170; right bottom subfigure highlights several highest recognition rates from left bottom subfigure. All four experiments are conducted on the CUFSF database.

Since the differences between NIR images and VIS images of the same subject are relatively small, we only utilize the RSLCR method to construct the final training set.

4.2 Experimental Settings

All the images used in this paper are aligned according to the eye centers. And the size of each image is cropped to 250×200. Each image patch is size of 10×10, and we keep 50% overlap between adjacent patches. All experiments are conducted on Windows 7 operation system with i7-4790 3.6G CPU, under the environment of MATLAB R2016b software. All the experimental results in this paper are the average of ten repetitions of the corresponding experiments by randomly partitioning the database.

Feature Exploration We explore four local feature descriptors, *i.e.* SIFT (Lowe 2004), HOG (Dalal and Triggs 2005), SURF (Bay, Tuytelaars, and Van Gool 2006), GABOR, and deep features *i.e.* VGG features (Parkhi, Vedaldi, and Zisserman 2015) in this paper. For local feature descriptors, we choose the default parameter settings. For deep features, we extract the output of last pooling layer in VGG-face networks. The left top subfigure in Figure. 5 gives the performance comparison with different features. It can be seen that deep features outperform other features. Hence, we represent images by deep features for the proposed approach in the following experiments.

In addition, the features utilized in this paper are dimension reduced by PCA. Thus, we explore the effectiveness of different feature dimensions. The accuracies corresponding to different dimensions of VGG features are shown in the right top subfigure in Figure. 5 and 750 is the best dimensionality for our framework.

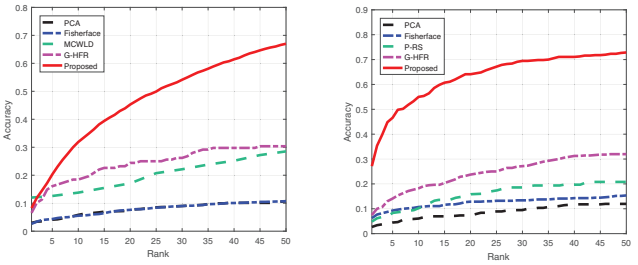

Figure 6: Left subfigure shows the results on IIIT-D database; right subfigure shows the results on Forensic Sketch database.

Discussions on the Combination of Different Face Sketch Synthesis Methods In this section, we choose eight synthesis methods (*i.e.* RSLCR, LLE, MRF, MWF, SFS, SRE, SVR and GANs) to explore the effectiveness of different combinations of synthesized image pairs for the proposed method. The synthesized examples can be found in Figure. 4. We select some number (from 1 to 8) of synthesized image pairs for joint training, and traverse all the combinations. Each experiment is repeated 10 times and the experimental results are shown in Figure. 5. The best four combinations are {RSLCR, MWF, SRE}, {RSLCR, SFS, SRE, GANs}, {RSLCR, SRE, GANs}, {RSLCR, MWF, GANs}. However, the SRE method has large computation complexity which costs too much time. Therefore, we choose the fourth best combination in all our following experiments. We also attempted other methods by incorporating different synthesized images into consideration. However, their improvements were marginal or even worse.

4.3 Results on Multiple Databases

For CUFSF database, we evaluate the rank-1 recognition accuracies under the same partition protocol. The comparison with the state-of-the-art methods is presented in Table 1. Our approach achieves the rank-1 recognition accuracy of 99.61%, which is superior to all other state-of-the-art methods.

For IIIT-D Sketch database and Forensic Sketch database, we compare the proposed approaches with state-of-the-art methods and the baseline methods. The experimental results are presented in Figure. 6. It can be seen that the proposed method outperforms state-of-the-art methods a lot. We list rank-50 accuracies in Table 2. They are 66.99% on IIIT-D

Algorithms	Rank-1 Recognition Accuracy
MRF(2009)	46.03%
MWF(2012)	74.15%
TFSPS(2013)	72.62%
RSLCR(2018)	75.94%
LRBP(2012)	91.12%
LDoGBP(2013)	91.04%
G-HFR(2017)	96.04%
PLS(2011)	51.00%
MvDA(2016)	55.50%
VGG(2015)	45.82%
SeetaFace(2017)	16.57%
JB(2017)+VGG(2015)	98.43%
JB(2017)+VGG(2015)+ Synthesis Methods	98.56%
AJL-HFR	**99.61%**

Table 1: Rank-1 recognition accuracies of the state-of-the-art approaches and our method on CUFSF database

Database	Algorithms	Rank-50 Recognition Accuracy
IIIT-D Sketch Database	MCWLD(2012)	28.52%
	G-HFR(2017)	30.36%
	AJL-HFR	**66.99%**
Forensic Sketch Database	P-RS(2013)	20.80%
	G-HFR(2017)	31.96%
	AJL-HFR	**72.86%**

Table 2: Rank-50 recognition accuracies of the state-of-the-art approaches and our method on IIIT-D database and Forensic Sketch database

Algorithms	Rank-1 Recognition Accuracy
LFDA(2011)	69.22%
CITE(2011	72.53%
LCKS-CSR(2012)	71.21%
P-RS(2013)	72.93%
CFDA(2014)	80.19%
VGG(2015)	62.91%
SeetaFace(2017)	69.50%
CEFD((2017))	83.93%
AJL-HFR	**99.05%**

Table 3: Rank-1 recognition accuracies of the state-of-the-art approaches and our method on CUHK VIS-NIR database

database and 72.86% on Forensic Sketch database for our proposed method in comparison to 30.36% and 31.96% of G-HFR which is the second best, *i.e.* we double the rank-50 accuracy of state-of-the-art methods.

The comparisons between the proposed method with state-of-the-art methods on CUHK VIS-NIR database are shown in Table 3. Our proposed approach achieve the rank-1 accuracy of 99.05%, which is much higher than the second best, 83.93% of (Gong et al. 2017).

5 Conclusion

In this paper we proposed an asymmetric joint learning method for HFR. The proposed AJL-HFR method jointly extracts intra-class information from the original training image pairs and the synthesized image pairs. The asymmetric framework is employed to avoid losing inter-class information. Experiments on viewed sketch (CUFSF) database, semi-forensic sketch (IIIT-D) database, forensic sketch database and near infrared image (CUHK VIS-NIR) database illustrate the effectiveness and superiority of the proposed method in comparison to state-of-the-art methods. In the future, we would explore more effective features besides VGG features.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (under Grants 61501339, 61772402, 61671339, U1605252), in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Natural Science Basic Research Plan in Shaanxi Province of China (under Grant 2017JM6085), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by CCF-Tencent Open Fund (under Grant IAGR 20170103), in part by the Fundamental Research Funds for the Central Universities under Grant JB160104, in part by the Program for Changjiang Scholars, in part by the Leading Talent of Technological Innovation of Ten-Thousands Talents Program under Grant CS31117200001, in part by the China Post-Doctoral Science Foundation under Grants 2015M580818 and 2016T90893, and in part by the Shaanxi Province Post-Doctoral Science Foundation.

References

Alex, A. T.; Asari, V. K.; and Mathew, A. 2013. Local difference of gaussian binary pattern: robust features for face sketch recognition. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, 1211–1216. IEEE.

Bay, H.; Tuytelaars, T.; and Van Gool, L. 2006. Surf: Speeded up robust features. *Computer vision–ECCV 2006* 404–417.

Belhumeur, P. N.; Hespanha, J. P.; and Kriegman, D. J. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence* 19(7):711–720.

Bhatt, H. S.; Bharadwaj, S.; Singh, R.; and Vatsa, M. 2012. Memetic approach for matching sketches with digital face images. Technical report.

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 *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 156–163. IEEE.

Chen, D.; Cao, X.; Wipf, D.; Wen, F.; and Sun, J. 2017. An efficient joint formulation for bayesian face verification.

IEEE Transactions on pattern analysis and machine intelligence 39(1):32–46.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, 886–893. IEEE.

Galoogahi, H. K., and Sim, T. 2012. Face sketch recognition by local radon binary pattern: Lrbp. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, 1837–1840. IEEE.

Gao, X.; Wang, N.; Tao, D.; and Li, X. 2012. Face sketch–photo synthesis and retrieval using sparse representation. *IEEE Transactions on circuits and systems for video technology* 22(8):1213–1226.

Gong, D.; Li, Z.; Huang, W.; Li, X.; and Tao, D. 2017. Heterogeneous face recognition: A common encoding feature discriminant approach. *IEEE Transactions on Image Processing* 26(5):2079–2089.

He, R.; Wu, X.; Sun, Z.; and Tan, T. 2017. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, 2000–2006.

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 07-49, University of Massachusetts, Amherst.

Ioffe, S. 2006. Probabilistic linear discriminant analysis. *Computer Vision–ECCV 2006* 531–542.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*.

Kan, M.; Shan, S.; Zhang, H.; Lao, S.; and Chen, X. 2016. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence* 38(1):188–194.

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

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

Lei, Z.; Zhou, C.; Yi, D.; Jain, A. K.; and Li, S. Z. 2012. An improved coupled spectral regression for heterogeneous face recognition. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, 7–12. IEEE.

Li, Z.; Gong, D.; Qiao, Y.; and Tao, D. 2014. Common feature discriminant analysis for matching infrared face images to optical face images. *IEEE Transactions on Image Processing* 23(6):2436–2445.

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

Liu, X.; Kan, M.; Wu, W.; Shan, S.; and Chen, X. 2017. Viplfacenet: an open source deep face recognition sdk. *Frontiers of Computer Science* 11(2):208–218.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.

Mittal, P.; Vatsa, M.; and Singh, R. 2015. Composite sketch recognition via deep network-a transfer learning approach. In *Biometrics (ICB), 2015 International Conference on*, 251–256. IEEE.

Parkhi, O. M.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition. In *BMVC*, volume 1, 6.

Peng, C.; Gao, X.; Wang, N.; and Li, J. 2017. Graphical representation for heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence* 39(2):301–312.

Sharma, A., and Jacobs, D. W. 2011. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 593–600. IEEE.

Song, Y.; Bao, L.; Yang, Q.; and Yang, M.-H. 2014. Real-time exemplar-based face sketch synthesis. In *European Conference on Computer Vision*, 800–813. Springer.

Susskind, J.; Hinton, G.; Memisevic, R.; and Pollefeys, M. 2011. Modeling the joint density of two images under a variety of transformations. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2793–2800. IEEE.

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

Wang, N.; Tao, D.; Gao, X.; Li, X.; and Li, J. 2013. Transductive face sketch-photo synthesis. *IEEE transactions on neural networks and learning systems* 24(9):1364–1376.

Wang, N.; Tao, D.; Gao, X.; and Li, X. 2014. A comprehensive survey to face hallucination. *International Journal of Computer Vision* 31(1):9–30.

Wang, N.; Gao, X.; and Li, J. 2018. Random sampling and locality constraint for face sketch. *Pattern Recognition* DOI: 10.1016/j.patcog.2017.11.008.

Zhang, J.; Wang, N.; Gao, X.; Tao, D.; and Li, X. 2011. Face sketch-photo synthesis based on support vector regression. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 1125–1128. IEEE.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 513–520. IEEE.

Zhou, H.; Kuang, Z.; and Wong, K.-Y. K. 2012. Markov weight fields for face sketch synthesis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 1091–1097. IEEE.