




Fine Tuning Dual Streams Deep Network with Multi-scale Pyramid Decision for Heterogeneous Face Recognition

Weipeng Hu¹ · Haifeng Hu¹ 

Published online: 26 October 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

In this paper, we propose a novel method called fine tuning dual streams deep network (FTDSDN) with multi-scale pyramid decision (MsPD) for solving heterogeneous face recognition task. As an extension of classical CNNs, FTDSDN can remove highly non-linear modality information and reserve the discriminative information using Rayleigh quotient objective function. Furthermore, we develop a powerful joint decision strategy called MsPD to adaptively adjust the weight of sub structure and obtain more robust classification performance. Experimental results show our proposed method achieves better performance on the challenging CASIA NIR-VIS 2.0 database, the heterogeneous face biometrics database, the CUHK face sketch FERET database, and the CUHK face sketch database, which demonstrates the effectiveness of our proposed approach.

Keywords Heterogeneous face recognition · Dual streams deep network · Multi-scale pyramid decision · Rayleigh quotient

1 Introduction

This paper focuses on the Heterogeneous Face Recognition (HFR) [9] matching problem, which has been widely studied in recent years. HFR task refers to matching a probe to the gallery taken from alternate imaging modality. The major difficulties of HFR lie in the great discrepancies between different image modalities, such as the identity related information, modality related information, face variations (e.g., illumination, poses, and expressions), etc.

During the last decade, many methods have been proposed to alleviate the appearance difference from heterogeneous data. Most of them can be generally categorized into four classes: synthesis-based model [19,34,37,40,48,49], coupled subspace learning [11,17,22,31,32,36], feature representation [1,5,16,25] and deep learning methods [7,24,29,30,38,42,47].

✉ Haifeng Hu
huhaif@mail.sysu.edu.cn

Weipeng Hu
huwp5@mail2.sysu.edu.cn

¹ School of Electronics and Information Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China

Synthesis-based methods [19,34,37,40,48,49] transform the heterogeneous face images to the same modality, and HFR task is converted into traditional face recognition problem, which is also called Face Hallucination (FH) and face sketch-photo synthesis techniques [37]. Song et al. [34] propose an effective sketch denoising method, in which Markov Random Field based methods can be formulated as the baseline improvements by adding smoothness constraints to reduce noise when synthesizing sketch patches. Li et al. [19] propose a learning-based framework synthesizing the normal face from the infrared input, which exploit the local linearity in both image spatial domain and image manifolds. Tang et al. [40] reduce the difference between photo and sketch by transforming a photo image into a sketch, which applies a Bayesian classifier to distinguish the probing sketch from the synthesized pseudo-sketches. However, the synthesis process is actually more difficult than recognition and the performance of these methods heavily depends on the fidelity of the synthesized images [28]. Coupled subspace learning based methods [11,17,22,31,32,36] map multimodal data into a common feature space to eliminate the large discrepancies of cross-modality image pairs. Lei et al. [17] propose Coupled Spectral Regression (CSR) method to coupling the cross-modality images into a discriminative subspace. Lin et al. [22] propose Common Discriminant Feature Extraction (CDFE) method for HFR matching, where two transforms are simultaneously learned to transform the samples in both modalities respectively to the common feature space. Sharma et al. [31] propose Partial Least Squares (PLS) to linearly map images in different modalities to a common linear subspace in which they are highly correlated. Tian et al. [36] adopt grassmannian Radial Basis Function (RBF) kernel to keep the relationship between subspaces, and use Kernel Canonical Correlation Analysis (KCCA) to handle correlation mapping between visible light (VIS) and near-infrared (NIR) domains. The main problem of coupled subspace learning methods is the projection procedure always causes information loss which may decrease the recognition performance [28]. Feature representation based methods [1,5,16,25] represent cross-modal face images by taking advantage of effective feature descriptors. The main purpose of these approaches is to reduce the cross-modality gap by exploring the most modality-insensitive features. Klare et al. [16] present Local Feature-based Discriminant Analysis (LFDA) to individually represent both sketches and photos. Based on the fact that the sketches are similar to their corresponding photos, Alex et al. [1] propose a face descriptor called Local Difference of Gaussian Binary Pattern (LDoGBP), which encodes the DoG representation of the image into a binary pattern. Lu et al. [25] propose a Coupled Simultaneous Local Binary Feature Learning and Encoding (CSLBFLE) which performs shared structured and latent feature learning to reduce the heterogeneous gap between face images of different modalities for heterogeneous face matching. Gong et al. [5] present a common encoding feature discriminant approach to reduce the modality gap at the feature extraction stage by converting the original face images pixel by pixel into a common encoded representation, and then infer the person's identity information for enhanced recognition performance. However, most existing methods represent an image ignoring the special spatial structure of faces, which is crucial for face recognition in reality [28].

Recently, many CNN-based methods are proposed for HFR task. Reale et al. [29] propose to extract extra information from a pre-trained visible face network and put forth an altered contrastive loss function to effectively train the network. Zhang et al. [47] conduct cross-modality conversion with Conditional Generative Adversarial Nets (cGAN), and further enhance the recognition performance by fusing multi-modal matching results. Wang et al. [38] introduce a 2D-3D HFR approach based on Deep Canonical Correlation Analysis (Deep CCA), which incorporates CNN into CCA, thus learning the mapping between hierarchically learned features of different modalities. Liu et al. [24] present a deep Transfer

NIR-VIS heterogeneous facE recognition neTwork (TRIVET) for solving HFR problem, which integrates the deep representation transferring and the triplet loss to get consolidated feature representations. Saxena et al. [30] pre-train and fine-tune the CNN model, and explore different metric learning strategies to reduce the discrepancies between the different modalities. By naturally combining subspace learning and invariant feature extraction into CNNs, He et al. [7] develop an invariant deep representation approach to map both NIR and VIS images to a compact Euclidean space. Wu et al. [42] propose a Coupled Deep Learning (CDL) approach by introducing low-rank relevance constraint and cross modal ranking into CNN. Though existing deep learning based methods lead to better performance in HFR matching, they have two limitations. Firstly, in most of the available CNNs [6,30], the softmax loss function is used as the supervision signal to train the deep model. In order to enhance the discriminative power of the deeply learned features, constructing a highly efficient loss function for discriminative feature learning in CNNs is non-trivial. Secondly, the individual features extracted by the above CNN-based methods are less discriminative and robust compared to the fusion of multiple deep features. To overcome these limitations, in this paper, we present Fine Tuning Dual Streams Deep Network (FTDSDN) with Multi-scale Pyramid Decision (MsPD) for HFR. Main contributions of our work can be summarized as follows:

- A novel supervised joint decision strategy MsPD, is presented to adaptively adjust the network weights according to the discriminating power of each sub network.
- An effective FTDSDN is developed to learn modality invariant representation, which avoids the network overfitting problem.
- Our FTDSDN employs Rayleigh quotient as objective function, which maps deep features into a discriminate feature space to decrease the intra-class variation while reserving the inter-class variation.
- Experimental results on the multiple challenging benchmark HFR datasets verify the effectiveness of the proposed model. In the following, we refer to Fine Tuning Dual Streams Deep Network (FTDSDN) with Multi-scale Pyramid Decision (MsPD) as F-MsPD.

The remainder of this paper is organized as follows. Section 2 presents the formulation of our F-MsPD model. Section 3 evaluates the performance of our method using two benchmark datasets. In Sect. 4, we conclude the paper.

2 The Proposed F-MsPD Approach

This section details the proposed F-MsPD model. As shown in Fig. 1, in our model, multiple parallel sub networks (i.e., FTDSDN) are used for discriminative feature extraction. Different from the two-stream ConvNet architecture [33], in which the spatial and temporal networks are completely independent without sharing parameters, the parameters of the dual streams deep network are partially shared in our FTDSDN. Then a new MsPD fusion strategy is implemented to adaptively adjust the network weights according to the discriminating performance of each sub network. And different from the Trunk-Branch Ensemble CNN model [2], which extracts complementary information from holistic face images and patches cropped around facial components, our F-MsPD model adaptively adjust the judgment weight of multiple parallel sub networks FTDSDN to make a joint decisions. Finally, we introduce our new proposed training strategy.

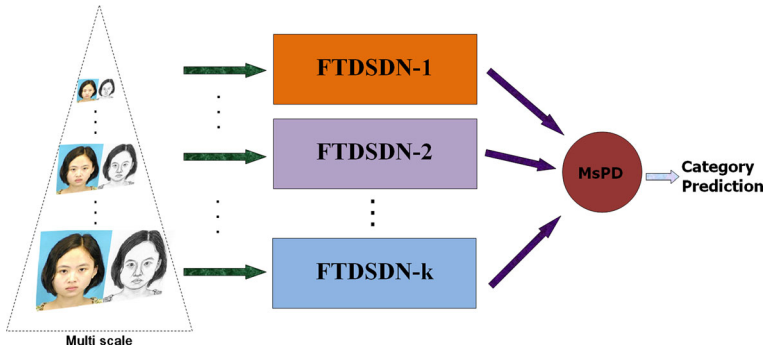


Fig. 1 Each scale pairs are processed by a corresponding sub structure FTDSN-*i* (*i* = 1, 2, ..., *k*, *k* representing the total number of scales.). And the final classification employs a newly proposed joint decision strategy MsPD

2.1 Multi-scale Pyramid Decision

In this section, MsPD, a novel fusion strategy for multiple independent unrelated sub networks, is presented to learn weight parameter adaptively for each FTDSN, which largely enhances HFR performance. In our model, for the *i*th scale the *m*th image I_{Nm}^i or I_{Vm}^i , we extract a feature vector x_{Nm}^i or x_{Vm}^i through FTDSN-*i*, where N and V respectively denote two modalities. We denote the final classification estimate as follows:

$$\widehat{\theta} = \arg \min \lambda_1 r_{\theta j}^1 + \lambda_2 r_{\theta j}^2 + \dots + \lambda_k r_{\theta j}^k \tag{1}$$

where λ_i (*i* = 1, 2, ..., *k*, *k* is the total number of the scale structure) is the weight of the *i*th FTDSN (denoted as FTDSN-*i*). For the θ th probe, the similarity ranking of the *j*th gallery is denoted as $r_{\theta j}^i$, where *i* indicates the *i*th scale structure FTDSN-*i*.

The Euclidean distance $D_{\theta j}^i$ between the probe $x_{N\theta}^i$ and all the galleries x_{Vj}^i (*j* = 1, 2, ..., *c*, where *c* is the total number of class) can be expressed as:

$$\begin{aligned} D_{\theta j}^i &= \langle X_{Nj}^i - X_{V\theta}^i, X_{Nj}^i - X_{V\theta}^i \rangle \\ D_{\theta}^i &= [D_{\theta 1}^i, D_{\theta 2}^i, \dots, D_{\theta c}^i] \end{aligned} \tag{2}$$

We sort D_{θ}^i and obtain r_{θ}^i with corresponding element position $r_{\theta i} = [r_{\theta 1}^i, r_{\theta 2}^i, \dots, r_{\theta c}^i]$, $r_{\theta j}^i \in \{1, 2, \dots, c\}$. $r_{\theta j}^i$ is called the similarity ranking for the probe. In order to achieve robust fusion, we adjust each $r_{\theta j}^i$ according to the threshold δ (our experimental settings $\delta = 10$)

$$r_{\theta j}^i = \begin{cases} r_{\theta j}^i & 1 \leq r_{\theta j}^i \leq \delta \\ \delta + 1 & r_{\theta j}^i > \delta \end{cases} \tag{3}$$

Two parameters η_i and β_i together determine the weight λ_i of MsPD, which can be formulated as below:

$$\lambda_i = \left(1 - \frac{\beta_i}{\sum_{r=1}^k \beta_r} \right) + \left(1 - \frac{\eta_i}{\sum_{r=1}^k \eta_r} \right) \tag{4}$$

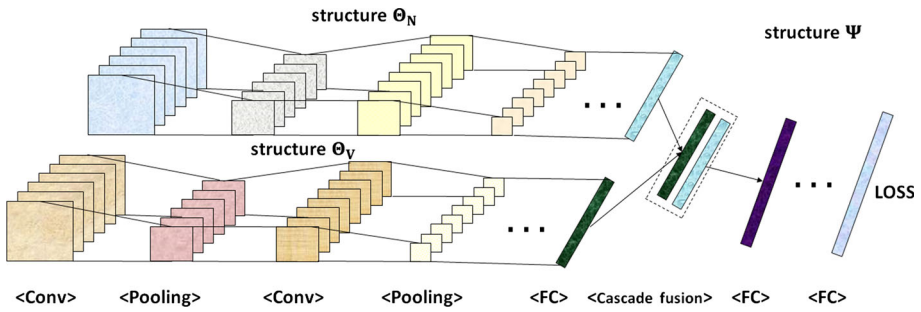


Fig. 2 The DSDN consists of separate sub structure Θ_q and a common sub structure Ψ

The parameter η_i describes the discriminative performance of the i th scale structure FTSDSN- i , which can be formulated as below:

$$\eta_i = \sum_{\theta=1}^n \sum_{j=1}^c \varepsilon(r_{\theta r}^i - r_{\theta j}^i) \tag{5}$$

where $r_{\theta r}^i$ is the similarity ranking of the label and $\varepsilon(\cdot)$ is the step function. In order to focus on easily misclassified examples and achieve better recognition performance, we design the parameter β_i to describe the ability of FTSDSN- i in dealing with these samples, which can be formulated as below:

$$\beta_i = \sum_{\theta=1}^n \sum_{j=1}^c \varepsilon(Q_{\theta r} - Q_{\theta j}) \varepsilon(r_{\theta r}^i - r_{\theta j}^i) \tag{6}$$

where

$$Q_{\theta j} = \beta_1 r_{\theta j}^1 + \beta_2 r_{\theta j}^2 + \dots + \beta_k r_{\theta j}^k \tag{7}$$

where $Q_{\theta r}$ is the joint probability ranking of the label. $\varepsilon(Q_{\theta r} - Q_{\theta j})$ adjudicates the hard, misclassified examples, while $\varepsilon(r_{\theta r}^i - r_{\theta j}^i)$ describes the ability of FTSDSN- i in dealing with hard examples. We assign an initial value $1/k$ to each β_i . Once we obtain weights distribution λ_i , we can determine the final output of F-MSPD via Eq. (1).

2.2 Dual Streams Deep Network

Inspired by the observations that removing highly non-linear modality information and reserving the discriminative information are useful for HFR task, we design Dual Streams Deep Network (DSDN) for feature extraction from two modal data.

Let I_{V_m} and I_{N_m} be the VIS and NIR images respectively, where $m = 1, 2, \dots, n$ denotes the m th sample. The DSDN feature extraction process is denoted as $x_{qm} = Conv(I_{qm}, \Theta_q, \Psi)$ ($q \in N, V$), where $Conv$ is the feature extraction function defined by the DSDN, x_{qm} is the extracted feature vector. The label space contains c unique classes, where each instance is associated with a corresponding label l_{qm} . As shown in Fig. 2, the DSDN consists of separate sub structure Θ_q and a common sub structure Ψ . Structure Θ_q eliminates the highly nonlinear modality information in the low-level features extraction stage, while

structure Ψ enhances discriminative and generalized identity information in the high-level features extraction stage. Note that the DSDN is designed with non-linear activation function [26], e.g. tanh, sigmoid, ReLU, PReLU and etc. Structure Θ_q and Ψ contain multiple convolutions, pooling layers and fully connected layers, which results in a deep architecture of DSDN.

The traditional CNN employs softmax as the cost function, in which the deeply learned features would contain large intra-class variations [41]. To extract discriminative features, the Fisher criterion [3] maximize distance between the classes and the minimize distance within the class. Inspired by Fisher criterion, in our work, we apply Rayleigh quotient [13] objective function to map deep features into a discriminate feature space to decrease the intra-class variation while reserving the inter-class variation, which contribute to reducing the gap between different modal domains. The Rayleigh quotient objective function can be expressed as follows

$$L(\Theta_q, \Psi) = \min \frac{Tr(S_W)}{Tr(S_B)} \quad (8)$$

where $Tr(\cdot)$ indicates the trace of the matrix, and S_W and S_B are the within- and between-class matrixes. Here, $S_W = \sum_{r=1}^c \sum_{q \in N, V} \sum_z \mathbb{I}(l_{qz} = r)(x_{qz} - m_r)(x_{qz} - m_r)^T$ and $S_B = \sum_{r=1}^c n_r(m_r - m)(m_r - m)^T$, where z is the index of sample in each class. $\mathbb{I}(\cdot)$ is the indicator function with value of 1 or 0. m_r is the mean value of the r th class with the number of n_r , and m is the mean value of all input samples with the number of n .

2.3 Network Structure

We design three different network structures F-MsPD to verify that the Rayleigh entropy objective function and the fusion strategy MsPD can be effectively applied to various deep network. And each F-MsPD model consists of three sub network FTSDSN- i ($i = 1, 2, 3$).

Two of the structures F-MsPD named F-MsPD-A and F-MsPD-B are shown in Table 1. In sub network FTSDSN-1 of F-MsPD-A model, both Θ_N^i and Θ_V^i are set to 2c-s-4c-s-360f, and Ψ is set to 80f, where lowercase letters c, s and f denote the convolution layer, mean pooling layer and fully connected layer respectively, while the digit represents the number of convolutional channels or neurons, and the kernel size of the convolution layer is set to 5×5 .

The parameter setting of the designed sub network FTSDSN- i in F-MsPD-C model is shown in Table 2, in which the three sub network FTSDSN- i ($i = 1, 2, 3$) have the same network structure. In sub network FTSDSN- i , Session 1–4 represent the Θ_N^i and Θ_V^i sub structure, and fonts have been bolded in Table 2, while Session 5–6 represent the Ψ sub structure. Compared to F-MsPD-A and F-MsPD-B, the F-MsPD-C model is wider and deeper, which is applicable to more complex HFR datasets, such as CASIA NIR-VIS 2.0 [20] datasets.

2.4 Optimization and Fine-Tuning Strategy

We employ the gradient descent following the chain rule to solve the optimization problem in DSDN. For the convolution parameters Θ_N^i and Ψ , we use conventional back-propagation method to update them. The optimization process of each scale structure is performed in batches of data. Specifically, we calculate the gradient of loss layer, i.e., the gradient of $L(\Theta_q, \Psi)$ with respect to x_{qm} . Similar to [13,43], S_W and S_B can be expressed as $S_W =$

Table 1 Parameter setting of the designed network in F-MsPD-A and F-MsPD-B model

Model	Structure	Input image scale	Θ_q^i	Ψ^i
F-MsPD-A	FTDSDN-1	32×32	2-s-4-s-360f	80f
	FTDSDN-2	72×72	2-s-3-s-360f	100f
	FTDSDN-3	100×100	2-s-3-s-360f	120f
F-MsPD-B	FTDSDN-1	32×32	5-s-10-s-360f	80f
	FTDSDN-2	72×72	5-s-10-s-360f	120f
	FTDSDN-3	100×100	5-s-10-s-360f	160f

$x A_W x^T$ and $S_B = x A_B x^T$ respectively, where $x = [x_{q1}, x_{q2}, \dots, x_{qb}]$. Here, $A_W = I - \sum_{k=1}^c \frac{1}{n_k} e^k e^{kT}$ and $A_B = (I - \frac{1}{n} e e^T) - A_W$. According to [4], the gradient $L(\Theta_q, \Psi)$ w.r.t. x_{qm} can be calculated as:

$$\frac{\partial L(\Theta_q, \Psi)}{\partial x_{qm}} = (-2A_B x^T (x A_B x^T)^{-1} (x A_W x^T) (x A_B x^T)^{-1} + 2A_W x^T (x A_B x^T)^{-1})^T e^m \tag{9}$$

Based on the chain rule, we compute the gradient of Ψ (i.e., the gradient of $L(\Theta_q, \Psi)$ w.r.t. Ψ) as:

$$\frac{\partial L(\Theta_q, \Psi)}{\partial \Psi} = \frac{\partial L(\Theta_q, \Psi)}{\partial x} \frac{\partial x}{\partial \Psi} \tag{10}$$

Similar to the process employed for networks Ψ , chain rule can be utilized to calculate the gradient of Θ_q (i.e., the gradient of $L(\Theta_q, \Psi)$ w.r.t. Θ_q):

$$\frac{\partial L(\Theta_q, \Psi)}{\partial \Theta_q} = \frac{\partial L(\Theta_q, \Psi)}{\partial x} \frac{\partial x}{\partial \Psi} \frac{\partial \Psi}{\partial \Theta_q} \tag{11}$$

To effectively train the DSDN, we present a novel fine-tuning strategy. In the early training epoch, we share the parameters of the separate sub structure Θ_N and Θ_V . And in the later training epoch, we alternately update the parameters of the Θ_N and Θ_V with mini-batches. The DSDN combines with Fine-Tuning strategy to form the FTDSDN.

2.5 Algorithm

The Algorithm of F-MsPD is summarized in Algorithm 1. We first randomly initialize the parameters Θ_q^i and Ψ^i of the i th scale structure FTDSDN- i . And then we pre train and fine tune the FTDSDN. Finally, the parameter η_i, β_i and λ_i of MsPD are obtained.

3 Experiments

In this section, a number of experiments are carried out on two biometric applications in support of the following two objectives:

Table 2 Parameter setting of the designed sub network FTSDSN-i in F-MSPD-C model

Structure	Name	Type	Filter Size	
Section 1	<i>Conv_1</i>	Convolution	$3 \times 3 \times 32$	
	BN_1	Batch normalization		
	PReLU_1	PReLU		
	<i>Conv_2</i>	Convolution	$3 \times 3 \times 32$	
	BN_2	Batch normalization		
	PReLU_2	PReLU		
Section 2	Pool_1	Max pooling	2×2	
	Section 2	<i>Conv_3</i>	Convolution	$3 \times 3 \times 64$
		BN_3	Batch normalization	
		PReLU_3	PReLU	
	<i>Conv_4</i>	Convolution	$3 \times 3 \times 64$	
	BN_4	Batch normalization		
PReLU_4	PReLU			
Section 3	Pool_2	Max pooling	2×2	
	Section 3	<i>Conv_5</i>	Convolution	$3 \times 3 \times 64$
		BN_5	Batch normalization	
		PReLU_5	PReLU	
	<i>Conv_6</i>	Convolution	$3 \times 3 \times 64$	
	BN_6	Batch normalization		
PReLU_6	PReLU			
Section 4	Pool_3	Max pooling	2×2	
	Section 4	<i>Conv_7</i>	Convolution	$3 \times 3 \times 128$
		BN_7	Batch normalization	
		PReLU_7	PReLU	
	<i>Conv_8</i>	Convolution	$3 \times 3 \times 128$	
	BN_8	Batch normalization		
PReLU_8	PReLU			
Section 5	Pool_4	Max pooling	2×2	
	Section 5	<i>Conv_9</i>	Convolution	$3 \times 3 \times 192$
		BN_9	Batch normalization	
		PReLU_9	PReLU	
	<i>Conv_10</i>	Convolution	$3 \times 3 \times 192$	
	BN_10	Batch normalization		
PReLU_10	PReLU			
Section 6	Pool_5	Max pooling	2×2	
	Section 6	<i>Conv_11</i>	Convolution	$3 \times 3 \times 256$
		BN_11	Batch normalization	
		PReLU_11	PReLU	
	Pool_6	Max pooling	2×2	
	Fc	Fully connected	128	
cost	Rayleigh quotient			

Algorithm 1 Iterative Algorithm for F-MsPD

Require: $I_N \in \mathbb{R}^{d \times n}$, $I_V \in \mathbb{R}^{d \times m}$, $I_N \in \mathbb{R}^n$, $I_V \in \mathbb{R}^m$. I_N, I_V sampled at multiple scales respectively and
 get: $I_N^i \in \mathbb{R}^{d_i \times n}$, $I_V^i \in \mathbb{R}^{d_i \times m}$, $I_N^i \in \mathbb{R}^n$, $I_V^i \in \mathbb{R}^m$.
Ensure:
 1: **Initialization** $\Theta_q^i, \Psi^i, i = 1, 2, \dots, k$.
 2: **Pre-training** FTSDSN-i
 3: **repeat**
 4: update (Θ_q^i, Ψ_i) using Eq. (9)–(11) with fixed other networks
 5: **until** reach the iteration number
 6: **Fine-tuning** FTSDSN-i
 7: **repeat**
 8: update (Θ_q^i, Ψ_i) using Eq. (9)–(11) with fixed other networks
 9: compute the loss $L(\Theta_q, \Psi)$ using Eq. (3)
 10: **until** Converges or reach the iteration number
 11: **Training MsPD**
 12: learn the parameter η_i using Eq. (5)
 13: learn the parameter β_i using Eq. (6)–(7)
 14: learn the parameter λ_i using Eq. (4)



Fig. 3 Example sketch-photo image pairs in the CUFS dataset. (Odd to photo images, even for the sketch images.). All the photo-sketch pairs are resized to three scales 32×32 (scale-1), 72×72 (scale-2) and 100×100 (scale-3)

- Investigate the various properties of the F-MsPD algorithm.
- Evaluate the F-MsPD algorithms on HFR problem by comparing performance with other proposed state-of-the-art methods such as Invariant Deep Representation (IDR) [7].

3.1 Datasets

CUFS dataset The CUHK Face Sketch (CUFS) dataset [48] is a public domain database which consists of the sketches and photos of 188 different persons. Each person has one photo and one sketch composed by artist. All of these face photos are taken at frontal view with a normal lighting condition and neutral expression. In order to form multi-scale images, all the photo-sketch pairs are resized to three scales 32×32 (scale-1), 72×72 (scale-2) and 100×100 (scale-3), and each scale image-pairs are processed by a scale structure, i.e. FTSDSN-1, FTSDSN-2 and FTSDSN-3. In our test, 88 photo-sketch pairs are applied as training data and the rest 100 pairs are applied as testing data. Some photo-sketch pair samples are shown in Fig. 3.



Fig. 4 Example sketch-photo image pairs in the CUFSS dataset. (Odd to photo images, even for the sketch images.). All the photo-sketch pairs are resized to three scales 64×64 (scale-1), 128×128 (scale-2) and 192×192 (scale-3)



Fig. 5 Examples NIR and VIS images of three different subjects from the HFB face database. (Odd to VIS images, even for the NIR images). All the NIR-VIS pairs are resized to three scales 32×32 (scale-1), 72×72 (scale-2) and 100×100 (scale-3)



Fig. 6 Examples NIR and VIS images of three different subjects from the CASIA NIR-VIS 2.0 Database. (Odd to VIS images, even for the NIR images). All the NIR-VIS pairs are resized to three scales 64×64 (scale-1), 128×128 (scale-2) and 192×192 (scale-3)

CUFSS dataset The CUHK Face Sketch FERET (CUFSS) dataset [40] is composed of 1194 individuals from FERET face database. Each person has one photo and one sketch which is drawn with shape exaggeration. In this experiment, 700 individuals of the

Table 3 We formed several F-MsPD versions by a combination of different techniques

Three sub network	Fusion method	External dataset	Network structure	Name
No	No	No	FTDSDN-1	FTDSDN-1
No	No	No	FTDSDN-2	FTDSDN-2
No	No	No	FTDSDN-3	FTDSDN-3
Yes	MsPD	No	F-MsPD-A	F-MsPD-A
Yes	MsPD	No	F-MsPD-B	F-MsPD-B
Yes	MsPD	CASIA WebFace	F-MsPD-A	F-MsPD-A-E
Yes	MsPD	CASIA WebFace	F-MsPD-B	F-MsPD-B-E
Yes	MsPD	CASIA WebFace	F-MsPD-C	F-MsPD-C-E

CUFSS database are randomly selected for training and the rest are used as the testing set. An average recognition rate is gained by repeating the experiment 20 random splits. In order to form multi-scale images, all the photo-sketch pairs are resized to three scales 64×64 (scale-1), 128×128 (scale-2) and 192×192 (scale-3), and each scale image-pairs are processed by a scale structure. Some photo-sketch pair samples are shown in Fig. 4.

HFB face database HFB Face Database [21] is collected with samples only from two views (i.e., visual image (VIS) and near infrared image (NIR)). There are totally 5097 images, including 2095 VIS and 3002 NIR from 202 persons in the database. All the NIR-VIS pairs are resized to three scales 32×32 (scale-1), 72×72 (scale-2) and 100×100 (scale-3). In protocol I, the training set contains 1062 VIS and 1487 NIR images from 202 subjects, where the rest constitutes the test set. Some NIR-to-VIS pair samples are shown in Fig. 5.

CASIA NIR-VIS 2.0 CASIA NIR-VIS 2.0 dataset [20] consists of 725 subjects in total. There are 1–22 VIS and 5–50 NIR face images per subject. Under the View 2 protocol, the evaluation is performed via the tenfold process and in each fold, 357 subjects are used for training while the remaining 358 subjects for testing. In order to form multi-scale images, all the NIR-VIS pairs are resized to three scales 64×64 (scale-1), 128×128 (scale-2) and 192×192 (scale-3). Some photo-sketch pair samples are shown in Fig. 6.

3.2 Training Methodology

In our experiments, we adopt two different training methodologies to train the F-MsPD model, one using the external dataset, and the other without using external dataset. We formed several F-MsPD versions by a combination of different techniques (number of sub networks, fusion method, external dataset and network structure) as shown in Table 3. When we use the external dataset, we first train each sub networks FTDSDN-*i* on the extra dataset CASIA WebFace [45], which contains 10,575 subjects and 494,414 images with softmax loss function. Then we further train each sub networks FTDSDN-*i* on the HFR dataset such as CASIA NIR-VIS 2.0 datasets according to different tasks with Rayleigh quotient objective function. When we do not use the external dataset, we only train each sub networks FTDSDN-*i* on the HFR dataset.

Table 4 Recognition rate of training samples (RRTS) for the three scale structure (FTDSDN-1, FTDSDN-2 and FTDSDN-3), and its corresponding learning parameter values (η_i^* , β_i^* and λ_i). The learning rate of FTDSDN-1, FTDSDN-2, FTDSDN-3 in group (a)–(c) is set to 0.01, 0.005, and 0.001 respectively, and the number of training iterations in group (a)–(c) is set to 200, 100, and 80 respectively

Group	Scale structure	RRTS(%)	η_i^*	β_i^*	λ_i
(a)	FTDSDN-3	100	1	0.97	1.97
	FTDSDN-2	98.86	0.29	0.46	0.75
	FTDSDN-1	98.86	0.71	0.57	1.28
(b)	FTDSDN-3	100	1	0.97	1.97
	FTDSDN-2	97.72	0.2	0.52	0.72
	FTDSDN-1	98.86	0.8	0.52	1.32
(c)	FTDSDN-3	93.18	0.04	0.11	0.15
	FTDSDN-2	98.86	0.97	0.92	1.89
	FTDSDN-1	100	1	0.97	1.97

3.3 Empirical Studies of the F-MsPD Properties

The following properties of F-MsPD are studied on the CUFS and HFB database: the effects of joint decision strategy MsPD and the proposed fine-tuning strategy. These experiments are performed on one random split of the database into training and testing samples.

3.3.1 The Effects of Multi-scale Pyramid Decision

The effects of Multi-scale Pyramid Decision: We evaluate the performance of F-MsPD compared with FTDSDN-1, FTDSDN-2, FTDSDN-3 and the simple equal weight (SEW) joint decision strategy (i.e. FTDSDN-SEW, the parameter λ_i of each sub network FTDSDN is all set to 1) to illustrate the effectiveness of the MsPD. In Table 4, group (a)–(c) list the recognition rate of training samples (RRTS) for FTDSDN-i adopting F-MsPD-A network structure in CUFS dataset, and its corresponding parameter values η_i^* , β_i^* and λ_i . The difference between the group (a)–(c) is the learning rate and the number of iterations for FTDSDN-i. As we can find, the parameters η_i^* , β_i^* and λ_i are positively correlated with the RRTS, which represents the performance of FTDSDN to a certain extent. This indicates our MsPD strategy can adaptively adjust the weights according to the discriminative performance of FTDSDN. Figure 7 show the recognition performance of FTDSDN-1, FTDSDN-2, FTDSDN-3, FTDSDN-SEW and F-MsPD, where group (a)–(c) differ in learning rate and the number of iterations for FTDSDN-i. From group (a)–(c) of Fig. 7, our F-MsPD outperforms FTDSDN-SEW and achieves better recognition accuracy. Especially in group (c), in which the performance of the three sub networks FTDSDN are quite different, and the performance of FTDSDN-SEW is even worse than a single network FTDSDN, while our F-MsPD can still achieves high recognition rate. Compared with FTDSDN-SEW, our joint decision strategy F-MsPD can adaptively adjust the weight of sub structure FTDSDN-i, and thus make a more robust decision, which shows that multiple FTDSDN with an effective fusion strategy MsPD aid HFR task.

To further verify the validity of the MsPD fusion method, we present the results of evaluations of five variants of our method in Table 5. As can be found, the recognition performance of F-MsPD-B and F-MsPD-A is always better than a single sub network FTDSDN-i with rank-1 accuracy 98.6% and 100% respectively. The F-MsPD-B-E method perform better than F-MsPD-B method in HFB dataset, with rank-1 accuracy of 99.7% and 98.6% respectively, indicating that training the network with additional dataset can further improve the

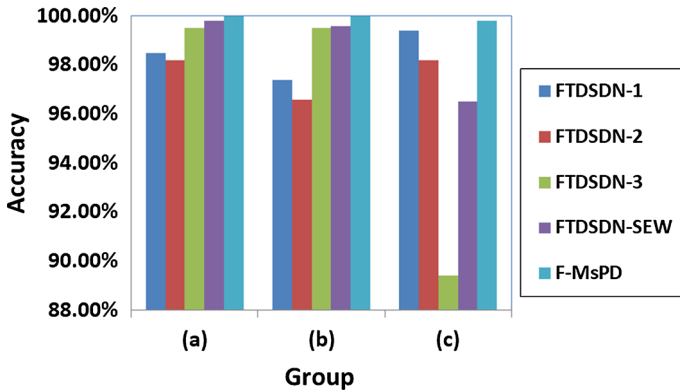


Fig. 7 Rank-1 accuracy on CUFS dataset. The learning rate of FTDSN-1, FTDSN-2, FTDSN-3 in group (a)–(c) is set to 0.01, 0.005, and 0.001 respectively, and the number of training iterations in group (a)–(c) is set to 200, 100, and 80 respectively

Table 5 Recognition result in CUFS and HFB database

Dataset	Method	Accuracy (%)
HFB	FTDSN-1	97.8
	FTDSN-2	97.9
	FTDSN-3	98.1
	F-MsPD-B	98.6
	F-MsPD-B-E	99.7
CUFS	FTDSN-1	99.4
	FTDSN-2	99.4
	FTDSN-3	99.5
	F-MsPD-A	100
	F-MsPD-A-E	100

Bold values are used to emphasize the best performance obtained by the listed methods

recognition performance. From the results, it is clear that MsPD fusion method and external dataset training both have a positive effect on the HFR task performance.

3.3.2 Evaluations of the Proposed Fine-Tuning Strategy

Evaluations of the proposed fine-tuning strategy: Fig. 8 shows the recognition performance of FTDSN-1, FTDSN-2 and FTDSN-3 adapting F-MsPD-A network structure with different pre-training iterations and proper fine-tuning iterations in CUFS dataset. As can be seen from the figure, without the pre-training step (i.e., the pre-training number is 0), FTDSN-i network has poor performance with accuracy only 96.5%. However, excessive pre training (the pre-training number is superfluous) will lead the network parameters deviates the optimal solution, and as a result the FTDSN fail to be micro adjustment by fine-tuning procedure. It shows that FTDSN-3, FTDSN-2 and FTDSN-1 achieve optimal generalization ability respectively in the number of iterations 80, 130 and 200 respectively, and the corresponding accuracy increase by 2.7%, 2.8% and 2.9% respectively compared to without fine-tuning strategy. Therefore, the pre-training and fine-tuning strategy can effectively train the whole network and improve the recognition performance of our FTDSN.

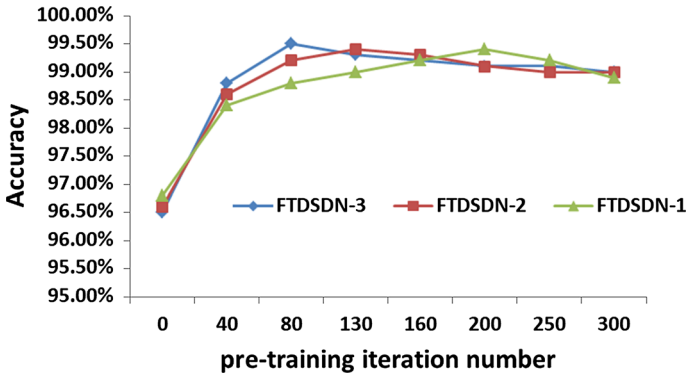


Fig. 8 Pre-training iteration number and the recognition rate of testing samples in the FTSDSN-1, FTSDSN-2 and FTSDSN-3 with proper fine-tuning

3.4 Comparison with State-of-the-Art Methods

In HFR experiments, we evaluate the F-MsPD algorithms on CASIA NIR-VIS 2.0, HFB, CUFSS and CUFS dataset, by comparing performance with other proposed state-of-the-art methods. The parameters setting for the compared algorithms are set according to the published papers [10,12].

CASIA NIR-VIS 2.0 dataset We compare the performance of F-MsPD-C-E with some approaches including Common Discriminant Feature Extraction (CDFE) [22], Multi-View Discriminant Analysis (MvDA) [14], VGG [27], SeetaFace [23], CenterLoss [41], and the state-of-the-art methods i.e., Yi et al. [44], Tian et al. [36], Lu et al. [25], COTS+Low-rank [18], Reale et al. [29] and IDR [7]. Table 6 shows rank-1 accuracy of different NIR-VIS face recognition methods. The IDR method achieves the best performance in rank-1 accuracy with 95.8%, which embeds two orthogonal subspaces into the deep network to play the role of orthogonal separation of modal information and spectral information, so as to extract discriminatory identity feature. Compared with the F-MsPD-C-E, the traditional HFR methods CDFE, MvDA, Tian et al. [36] and Lu et al. [25] have lower rank-1 accuracy, whose accuracy is 27.9%, 41.6%, 82.6% and 86.9% respectively, which shows that the deep feature extracted by F-MsPD-C-E is more discriminant than the traditional method, as deep methods can remove the modal information and retain more identity information at the same time. Compared with other deep learning methods such as VGG, SeetaFace, Yi et al. [44], CenterLoss, COTS+Low-rank, and Reale et al. [29], our F-MsPD-C-E has the highest 93.5% rank-1 accuracy, mainly because we use Rayleigh quotient loss function to effectively train the network. In addition, our model takes full advantage of the multiple sub networks and achieves a better classification performance using an effective fusion strategy MsPD.

HFB face database We perform comparison of F-MsPD-B-E with existing approaches PCA [15], Canonical Correlational Analysis (CCA) [46], CDFE [22], Linear Coupled Spectral Regression (LCSR) [17], Kernel Coupled Spectral Regression (KCSR) [17], Linear Discriminative Spectral Regression (LDSR) [10], and the state-of-the-art methods including Yi et al. [44] and Reale et al. [29] on HFB database. As shown in Table 7, the performance of our F-MsPD-B-E model is better than the traditional method PCA, CDFE, LCSR, LDSR et al., indicating the deep feature extracted by F-MsPD-B-E is more discriminative than the traditional method with rank-1 accuracy 99.7%. Our F-MsPD-B-E achieves an improvement up to 0.2% and 0.3% respectively compared to deep learning method Reale et al. [29] and

Table 6 Recognition rates in the CASIA NIR-VIS 2.0 dataset

Method	Accuracy (%)
CDFE [22]	27.9
MvDA [14]	41.6
VGG [27]	62.1
SeetaFace [23]	68.0
Tian et al. [36]	82.6
Yi et al. [44]	86.2
Lu et al. [25]	86.9
CenterLoss [41]	87.7
COTS+low-rank [18]	89.6
Reale et al. [29]	92.6
IDR [7]	95.8
F-MsPD-C-E	93.5

Bold values are used to emphasize the best performance obtained by the listed methods

Table 7 Recognition result in HFB database

Method	Accuracy (%)
PCA [15]	12.1
CDFE [22]	97.21
CCA [46]	95.4
LCSR [17]	97.5
KCSR [17]	97.3
LDSR [10]	97.5
Yi et al. [44]	99.4
Reale et al. [29]	99.5
F-MsPD-B-E	99.7

Bold values are used to emphasize the best performance obtained by the listed methods

Yi et al. [44], indicating that our carefully designed structure FTSDSN and multi-scale joint decision strategy MsPD are both benefit for HFR task. FTSDSN structure can remove highly non-linear modality information and reserve the discriminative information with Rayleigh quotient objective function by exploiting the correlations from both inter- and intra-class data, and the proposed MsPD shows fused features is better than individual features for HFR task.

CUFSS dataset Table 8 shows the rank-1 accuracy of different sketch-photo methods including CCA [46], Partial Least Squares (PLS) [31], CDFE [22], MvDA [14], and the state-of-the-art methods i.e., Large Margin Coupled Feature Learning (LMCFL) [12] and Yi et al. [44]. The traditional methods CCA, PLS, CDFE, MvDA and LMCFL performed poorly in HFR tasks compared to the F-MsPD-C-E, with rank-1 accuracy of 38.4%, 48.1%, 50.4%, 56.2% and 80.5% respectively, indicating that those traditional methods are difficult to extract robust features for HFR task. Yi et al. [44] performs worse than our F-MsPD-C-E, because it extracts Gabor features as input of the Restricted Boltzmann Machines, which may lead to loss of identity information and introduce modal information at the same time. Our F-MsPD-C-E has the best performance at rank-1 accuracy with 99.1% compare to the

Table 8 Recognition rates in the CUFSF dataset for sketch-photo face recognition

Method	Accuracy (%)
CCA [46]	38.4
PLS [31]	48.1
CDFE [22]	50.4
MvDA [14]	56.2
LMCFL [12]	80.5
Yi et al. [44]	98.6
F-MsPD-C-E	99.1

Bold values are used to emphasize the best performance obtained by the listed methods

Table 9 Recognition rates in the CUFS dataset for sketch-photo face recognition

Method	Accuracy (%)
KCSR [17]	83.0
PLS [31]	93.6
BLM [35]	94.2
CCA [46]	94.6
LDSR [10]	95.0
SCDL [39]	95.2
PCA-STL [8]	97.4
Yi et al. [44]	100
F-MsPD-A-E	100

Bold values are used to emphasize the best performance obtained by the listed methods

state-of-the-art methods, which shows that the F-MsPD-C-E method can effectively train the network by using the Rayleigh quotient loss function. In addition, to make full use of the multiple sub structure FTSDN-i, our MsPD fusion method adaptively adjust the weight of sub structure and obtain more robust classification performance. Therefore, our F-MsPD method is more suitable for HFR tasks.

CUFS dataset We compare the performance of the proposed algorithms against the state-of-the-art approaches including KCSR [17], PLS [31], Bilinear Model (BLM) [35], CCA [46], LDSR [10], Semi-coupled Dictionary Learning (SCDL) [39], PCA+STL [8] and the state-of-the-art methods i.e., Yi et al. [44]. As shown in Table 9, the recognition rates of PLS and CCA are only 93.6% and 94.6% respectively, mainly because they are both linear methods, which lack the robustness for high nonlinearity of heterogeneous face images. BLM, LDSR, KCSR, SCDL, PCA+STL have poorer performances mainly because they cannot extract discriminative features to fit for the untrained data well. Our F-MsPD-A-E achieves the highest recognition performance with accuracy 100%. All of these results suggest that the features learned by FTSDN are not only separable but also discriminative with Rayleigh quotient objective function, which proves that multiple FTSDN with an effective fusion strategy MsPD aid HFR task.

3.5 Discussions

We have performed a large number of experiments on HFR task to evaluate our proposed algorithms. From the results presented above, the following observations are made.

- The F-MsPD achieves the best accuracy compared with single networks FTSDSN, indicating that our joint decision strategy MsPD can adaptively adjust the network weights according to the discriminating performance of each sub network, which shows fused features is better than individual features for HFR matching.
- FTSDSN-3, FTSDSN-2 and FTSDSN-1 network adopting fine-tuning strategy can further improve recognition performance, and the corresponding accuracy increase by 2.7%, 2.8% and 2.9% respectively compared to without fine-tuning strategy, which shows that fine-tuning strategy can effectively train the network.
- The designed three model F-MsPD-A-E, F-MsPD-B-E, and F-MsPD-C-E are effective for HFR task, which shows that the Rayleigh entropy objective function and the fusion strategy MsPD can be effectively applied to various deep network structures.
- Extensive experiments on the benchmark heterogeneous face databases indicate the effectiveness of our proposed approach. This comparative evaluation demonstrates that F-MsPD is a robust and effective algorithm for HFR problem.

4 Conclusion

In this paper, we have developed a novel Fine Tuning Dual Streams Deep Network with Multi-scale Pyramid Decision for dealing with HFR problem. A novel supervised joint decision strategy MsPD, which shows fused features is better than individual features for HFR matching, is presented to adaptively adjust the network weights according to the discriminating performance of each sub network. Different from existing methods, our FTSDSN can exploit the correlations contained in both inter-view and intra-view data. Moreover, a novel fine tuning strategy is defined to effectively train the whole network. Experiments on four HFR datasets demonstrate the superior of our method over existing technique.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grants 61673402, 61273270, and 60802069, in part by the Natural Science Foundation of Guangdong under Grants 2017A030311029, 2016B010109002, 2015B090912001, 2016B010123005, and 2017B090909005, in part by the Science and Technology Program of Guangzhou under Grants 201704020180 and 201604020024, and in part by the Fundamental Research Funds for the Central Universities of China.

References

1. Alex AT, Asari VK, Mathew A (2013) Local difference of Gaussian binary pattern: robust features for face sketch recognition. In: IEEE international conference on systems, man, and cybernetics, pp 1211–1216
2. Ding C, Tao D (2018) Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Trans Pattern Anal Mach Intell* 40(4):1002–1014
3. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Hum Genet* 7(2):179–188
4. Fukunaga K (2013) Introduction to statistical pattern recognition. Academic press, Cambridge
5. Gong D, Li Z, Huang W, Li X, Tao D (2017) Heterogeneous face recognition: a common encoding feature discriminant approach. *IEEE Trans Image Process* 26(5):2079–2089
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference computer vision and pattern recognition, pp 770–778
7. He R, Wu X, Sun Z, Tan T (2017) Learning invariant deep representation for NIR-VIS face recognition. In: AAAI conference on artificial intelligence, vol 4, p 7
8. Hou CA, Yang M, Wang YCF (2014) Domain adaptive self-taught learning for heterogeneous face recognition. In: International conference on pattern recognition. IEEE, pp 3068–3073

9. Hu S, Short N, Riggan BS, Chasse M, Sarfraz MS (2017) Heterogeneous face recognition: recent advances in infrared-to-visible matching. In: IEEE international conference on automatic face and gesture recognition, pp 883–890
10. Huang X, Lei Z, Fan M, Wang X, Li SZ (2013) Regularized discriminative spectral regression method for heterogeneous face matching. *IEEE Trans Image Process* 22(1):353–362
11. Huo J, Gao Y, Shi Y, Yang W, Yin H (2018) Heterogeneous face recognition by margin-based cross-modality metric learning. *IEEE Trans Cybern* 48(6):1814–1826
12. Jin Y, Lu J, Ruan Q (2015) Large margin coupled feature learning for cross-modal face recognition. In: International conference on biometrics. IEEE, pp 286–292
13. Kan M, Shan S, Chen X (2016) Multi-view deep network for cross-view classification. In: IEEE conference computer vision and pattern recognition, pp 4847–4855
14. Kan M, Shan S, Zhang H, Lao S, Chen X (2016) Multi-view discriminant analysis. *IEEE Trans Pattern Anal Mach Intell* 38(1):188–194
15. Karlpearson FRS (1901) LIII. On lines and planes of closest fit to systems of points in space. *Philos Mag* 2(11):559–572
16. Klare BF, Li Z, Jain AK (2011) Matching forensic sketches to mug shot photos. *IEEE Trans Pattern Anal Mach Intell* 33(3):639
17. Lei Z, Li SZ (2009) Coupled spectral regression for matching heterogeneous faces. In: IEEE conference computer vision and pattern recognition, pp 1123–1128
18. Lezama J, Qiu Q, Sapiro G (2017) Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding. In: IEEE conference computer vision and pattern recognition, pp 6807–6816
19. Li J, Hao P, Zhang C, Dou M (2008) Hallucinating faces from thermal infrared images. In: IEEE international conference on image processing, pp 465–468
20. Li S, Yi D, Lei Z, Liao S (2013) The CASIA NIR-VIS 2.0 face database. In: IEEE conference computer vision and pattern recognition workshops, pp 348–353
21. Li SZ, Lei Z, Ao M (2009) The HFB face database for heterogeneous face biometrics research. In: IEEE conference computer vision and pattern recognition, pp 1–8
22. Lin D, Tang X (2006) Inter-modality face recognition. In: European conference on computer vision, pp 13–26
23. Liu X, Kan M, Wu W, Shan S, Chen X (2017) VIPLFaceNet: an open source deep face recognition SDK. *Front Comput Sci* 11(2):208–218
24. Liu X, Song L, Wu X, Tan T (2016) Transferring deep representation for NIR-VIS heterogeneous face recognition. In: International conference on biometrics, pp 1–8
25. Lu J, Erin LV, Zhou J (2017) Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. *IEEE Trans Pattern Anal Mach Intell* PP(99):1–1
26. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: International conference on international conference on machine learning, pp 807–814
27. Parkhi OM, Vedaldi A, Zisserman A, et al (2015) Deep face recognition. In: British machine vision conference, vol 1, p 6
28. Peng C, Gao X, Wang N, Li J (2017) Graphical representation for heterogeneous face recognition. *IEEE Trans Pattern Anal Mach Intell* 39(2):301–312
29. Reale C, Lee H, Kwon H (2017) Deep heterogeneous face recognition networks based on cross-modal distillation and an equitable distance metric. In: IEEE conference computer vision and pattern recognition workshops, pp 32–38
30. Saxena S, Verbeek J (2016) Heterogeneous face recognition with CNNs. In: European conference on computer vision, pp 483–491
31. Sharma A, Jacobs DW (2011) Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In: IEEE conference computer vision and pattern recognition, pp 593–600
32. Shi H, Wang X, Yi D, Lei Z, Zhu X, Li SZ (2017) Cross-modality face recognition via heterogeneous joint bayesian. *IEEE Signal Process Lett* 24(1):81–85
33. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp 568–576
34. Song Y, Bao L, Yang Q, Yang MH (2014) Real-time exemplar-based face sketch synthesis. In: European conference on computer vision. Springer, pp 800–813
35. Tenenbaum J (2000) Separating style and content with bilinear models. *Neural Comput* 12:1247–1283
36. Tian Y, Yan C, Bai X, Zhou J (2017) Heterogeneous face recognition via Grassmannian based nearest subspace search. In: IEEE international conference on image processing, pp 1077–1081
37. Wang N, Tao D, Gao X, Li X, Li J (2014) A comprehensive survey to face hallucination. *Int J Comput Vis* 106(1):9–30

38. Wang S, Huang D, Wang Y, Tang Y (2017) 2D–3D heterogeneous face recognition based on deep canonical correlation analysis. In: Chinese conference on biometric recognition. Springer, pp 77–85
39. Wang S, Zhang L, Liang Y, Pan Q (2012) Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: IEEE conference computer vision and pattern recognition, pp 2216–2223
40. Wang X, Tang X (2009) Face photo-sketch synthesis and recognition. *IEEE Trans Pattern Anal Mach Intell* 31(11):1955–1967
41. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. Springer, pp 499–515
42. Wu X, Song L, He R, Tan T (2017) Coupled deep learning for heterogeneous face recognition. arXiv preprint [arXiv:1704.02450](https://arxiv.org/abs/1704.02450)
43. Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S (2006) Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans Pattern Anal Mach Intell* 29(1):40
44. Yi D, Lei Z, Li SZ (2015) Shared representation learning for heterogenous face recognition. In: IEEE international conference and workshops on automatic face and gesture recognition, vol 1, pp 1–7
45. Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. arXiv preprint [arXiv:1411.7923](https://arxiv.org/abs/1411.7923)
46. Yi D, Liu R, Chu R, Lei Z, Li SZ (2007) Face matching between near infrared and visible light images. In: International conference on biometrics. Springer, pp 523–530
47. Zhang W, Shu Z, Samaras D, Chen L (2017) Improving heterogeneous face recognition with conditional adversarial networks. arXiv preprint [arXiv:1709.02848](https://arxiv.org/abs/1709.02848)
48. Zhang W, Wang X, Tang X (2011) Coupled information-theoretic encoding for face photo-sketch recognition. In: IEEE conference computer vision and pattern recognition, pp 513–520
49. Zhong J, Gao X, Tian C (2007) Face sketch synthesis using E-HMM and selective ensemble. In: IEEE international conference on acoustics, speech and signal processing, pp 485–488

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.