2D-3D Heterogeneous Face Recognition based on Deep Coupled Spectral Regression

Yangtao Zheng, Di Huang, Weixin Li, Shupeng Wang, Yunhong Wang
IRIP Lab, School of Computer Science and Engineering,
Beihang University, Beijing 100191, China
{ytzheng, dhuang, weixinli, wangshupeng, yhwang}@buaa.edu.cn

Abstract

As one of the major branches in Face Recognition (FR), 2D-3D Heterogeneous FR (HFR), where face comparison is achieved across the texture and shape modalities, has become more important. This paper proposes a novel deep learning based end-to-end approach, namely Deep Coupled Spectral Regression (DCSR), for such an issue. It jointly makes use of both the advantages of CNN based deep features and CSR based common subspace. Specifically, from 2D texture and 3D depth face maps, DCSR extracts more powerful features by a deep network with the cross-modality triplet loss, which show much better uniqueness and robustness than the hand-crafted ones. Further, DCSR learns the shared space between different modalities with the constraints of sample labels, and is thereby more discriminative than the widely used unsupervised methods. More importantly, the two steps above are integrated through a couple layer to explicitly optimize the weights of deep features and projection directions rather than a simple combination. Experiments are carried out on the FRGC v2.0 database, and the results reported clearly demonstrate the competency of our proposed method. Its generalization ability is also validated by additional experiments conducted on the CASIA NIR-VIS 2.0 database.

1. Introduction

Face Recognition (FR) is one of the most popular topics in the domain of computer vision and pattern recognition. The last several decades have witnessed its large progress. Traditional FR scenarios, *e.g.* 2D-2D or 3D-3D FR [11, 6, 15, 27], demand gallery and probe data in the same modality, *i.e.*, 2D texture and 3D shape, or even require the data captured by the same type of sensing devices. In contrast to them, Heterogeneous FR (HFR) matches face images of different modalities [19], such as visible lighting faces *vs.* near infrared faces [4], face photos *vs.* face sketches [29], and face texture maps *vs.* face shape maps. It has received increasing attention due to its scientific challenges and application potentials. Recently, along with the popularization of 3D FR using portable depth cameras (*e.g.* unlocking at iPhone), 2D-3D HFR has been investigated more extensively, which aims to provide a solution to FR where different views of faces are available in the 2D texture and 3D shape modality respectively. It becomes more useful since various new 2D, 3D, and RGB-D sensors are emerging, and plays a crucial role in retrieving faces across data for biometric, forensic, and entertainment systems. See Figure 1 for an illustration of 2D-3D HFR in the scenario of identification (1:N) and verification (1:1).

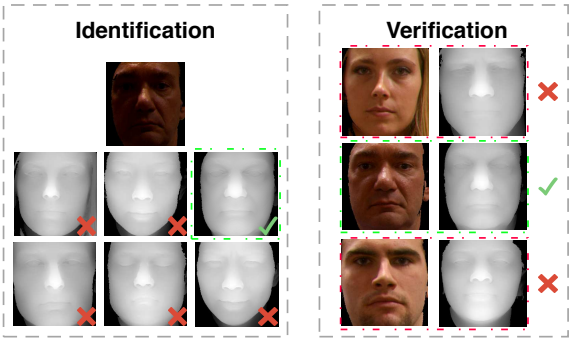


Figure 1. Illustration of 2D-3D HFR in the scenario of identification (left) and verification (right). The green boxes denote correct decisions.

In HFR, there exist two major issues, *i.e.* facial representation in different modalities and mapping learning between different facial features. For the former, in 2D-3D HFR, a number of features have been attempted, including holistic ones, *e.g.* Principal Component Analysis (PCA) coefficients produced on original pixels [5] and local ones, *e.g.* Local Binary Patterns (LBP) [9, 8] and Oriented Gradient Maps (OGM) [10, 7]. Wang *et al.* substantially improve this step by a deep model [24], namely Convolutional Neural Networks (CNNs), which hierarchically builds deep fa-

cial representations on both 2D texture maps and 3D shape maps. The deep features prove more effective compared with the previous hand-crafted ones and present high tolerance to face alignment. For the latter, subspace learning techniques are usually exploited to generate the common space across 2D and 3D face maps, and Canonical Correlation Analysis (CCA) [5, 8, 7, 24] as well as its kernel version [26] are major representatives. Current methods report promising performance; however, they generally work in a person dependent manner where the identity in the test set is required to enroll in the training phase, which degrades the generalization ability to unseen persons. Lei and Li [13] propose Coupled Spectral Regression (CSR) for HFR between facial images acquired under visible and near infrared lightings, and they adopt a supervised approach with a similar idea to Linear Discriminant Analysis (LDA) where inter-class distances are increased and intra-class ones are reduced, thus leading to a more discriminative shared subspace than that in the CCA family.

In this paper, we propose a novel deep learning approach to 2D-3D HFR, called Deep Coupled Spectral Regression (DCSR). It jointly makes use of both the advantages of CNN based deep features and CSR based common subspace. Specifically, DCSR extracts more powerful features from 2D texture and 3D depth face maps by a deep CNN model with the cross-modality triplet loss, which show much better uniqueness and robustness than the hand-crafted ones. Meanwhile, DCSR sufficiently exploits labels of samples, and the common space learned between different modalities is thus more discriminative and better generalized than Deep CCA [24]. More importantly, DCSR is an end-to-end model, integrating the two steps above through a couple layer to explicitly optimize the weights in deep features and projection directions rather than a simple combination of CNN and CSR. Extensive experiments are conducted on the FRGC v2.0 database, and the results are state-of-the-art, demonstrating the effectiveness of the proposed method. In addition, the results reported on CASIA NIR-VIS 2.0 show that the proposed method has a good generalization ability to other HFR problems.

In summary, the contributions are three-fold as follows:

- A new end-to-end supervised deep learning approach, namely DCSR, is proposed for 2D-3D HFR.
- For joint optimization in feature extraction and mapping learning, a novel layer structure, called couple layer, is designed, which combines CNN and CSR effectively and efficiently.
- State-of-the-art results are reached on the FRGC v2.0 database in 2D-3D HFR, and the scores of NIR-VIS HFR achieved on the CASIA NIR-VIS 2.0 database are also comparable to the best ones so far reported.

The rest of the paper is organized as follows. Section 2 briefly reviews the most related work of 2D-3D HFR in the literature. Section 3 introduces the proposed Deep Coupled Spectral Regression (DCSR) in detail, and Section 4 displays and analyzes the experimental results. Section 5 concludes the paper.

2. Related Work

To the best of our knowledge, the first attempt on 2D and 3D face matching dates back to 2005. Riccio and Dugelay claim that facial geometry is intrinsic and remains stable in 2D and 3D data, and with the help of a set of pre-defined fiducial points, they calculate several geometrical invariants to associate facial texture and shape images [22]. Promising results are delivered on a small database with 50 people, but when the number of identities greatly enlarges, the discriminability of this feature tends to be problematic. Furthermore, precisely localizing those landmarks on both 2D and 3D faces, especially in the wild, is itself a difficult task.

Rama *et al.* [21] present a 2D-3D HFR method, where 3D data used for training are cylindrical (180° in the yaw axis) texture images of whole faces rather than their depth maps and ordinary 2D facial images are employed for test. Partial Principal Component Analysis (P^2CA) is applied for feature extraction in a low dimensional subspace. An accuracy of more than 90% is reached on a dataset of 18 persons, even if pose variations occur (in the yaw direction). However, the claimed 3D data actually offer texture clues, therefore, this method is more sensitive to lighting changes.

In [26], Yang *et al.* exploit CCA to generate the mapping between the 2D texture and 3D depth maps of faces. Instead of applying CCA on facial images globally, a patch based strategy is utilized to divide faces into some uniform blocks, and CCA operates on individual pairs of patches of the two modalities and their contributions are then combined for decision making. They further enhance this approach by Kernel CCA (KCCA), and a result around of 85% is reported on the samples of 28 subjects. Unlike the previous studies, the experiment is carried out in a person independent way. Nevertheless, raw pixel based features are often criticized for its robustness, which leaves much space for improvement.

Huang *et al.* [5, 8] regard textured 3D face models as gallery samples and 2D face texture images as probe ones, and build the framework consisting of two separate matching processes, *i.e.* 2D-2D FR and 2D-3D HFR, whose similarity scores are fused for prediction. In their 2D-3D HFR phase, they incorporate LBP facial features into CCA to improve the performance in the presence of lighting changes. In the subsequent, they replace LBP features with the biological vision inspired feature, namely OGM, for performance gain [7]. Their method achieves state of the art result on the comprehensive FRGC v2.0 dataset but under the person dependent setting. Additionally, it requires sophisti-

cated preprocessing on illumination and pose variations.

Zhang *et al.* [28] proposed a framework, which combines the generative capacity of Conditional Generative Adversarial Nets (cGAN) and the discriminative power of deep CNN features for cross-modality learning. Firstly, they conduct 3D depth image reconstruction from a single 2D texture image with cGAN, and the recovered depth information enables a straightforward comparison in the 3D space. To extract features of different modalities and provide pre-trained models for cross-modality learning, two discriminative CNN models are trained individually. They further enhance the recognition performance by fusing multi-modal matching results. While attractive, this method highly depends on the quality of the reconstructed 3D depth map, especially with the same identity information preserved in the 2D image.

Wang *et al.* [24] introduce Deep CCA in 2D-3D HFR, where CNN based deep features are computed on both the 2D texture and 3D shape maps for facial representation. The hierarchically learned features are expected to be more discriminative and less affected by the changes in illumination and head pose. When the deep features are extracted, they are fed into CCA simultaneously, and the correlation gradient is calculated to optimize the deep neural network. This framework indicates that the deep features are superior to hand-crafted ones, while similar to [7], the good results are only produced in person dependent experiments.

Despite great progress made in 2D-3D HFR as the increasing performance shows, the conventional methods basically use hand-crafted features, which limit both the accuracy and robustness. Deep features prove more effective, but the succeeding mapping is learned in an unsupervised way, leaving room for improvement. Meanwhile, they generally work in a person dependent manner, where identities in the gallery and probe sets overlap with each other. Under this assumption, they cannot work well in the real world (*e.g.* the open-set recognition scenario), which requires a strong generalization ability. Compared to them, the proposed method adopts a DCSR framework, which represents 2D and 3D faces more comprehensively and learns the heterogeneous mapping more accurately.

3. Deep Model for 2D-3D HFR

In this section, we present our deep common subspace learning framework for 2D-3D HFR. We first briefly revisit the original CSR method, and then describe our proposed DCSR in detail.

3.1. Coupled Spectral Regression

The Coupled Spectral Regression (CSR) method, proposed by Lei and Li [13], aims at generating a projection which can map heterogenous data (*i.e.*, visible lighting faces and near infrared faces) into a common subspace, and it proves more discriminative for face identification and verification than the CCA related methods, as it makes use of sample labels for supervision.

Formally, given two heterogeneous sample sets $\{X^1, X^2\}$, where $X^1 = [x_1^1, \ldots, x_n^1]$, $X^2 = [x_1^2, \ldots, x_n^2]$, and n denotes the size of the sample set, we generate their low-embeddings $\{Y^1, Y^2\}$, where $Y^1 = [y_1^1, \ldots, y_n^1]^T$, $Y^2 = [y_1^2, \ldots, y_n^2]^T$, and $Y^1, Y^2 \in \mathbb{R}^{n \times d}$. With linear assumption, the low-embeddings can be defined as

$$Y^1 = {X^1}^T A^1, Y^2 = {X^2}^T A^2, \tag{1}$$

where A^1 and A^2 are the projection matrices for X^1 and X^2 respectively.

The objective function of CSR is formulated as

$$\begin{aligned} \{A^1, A^2\} = \underset{A^1, A^2}{\operatorname{argmin}} \{ & \frac{1}{n} \|Y^1 - {X^1}^T A^1\|^2 \\ & + \frac{1}{n} \|Y^2 - {X^2}^T A^2\|^2 + \eta \|A^1 - A^2\|^2 \\ & + \lambda(\|A^1\|^2 + \|A^2\|^2)\}. \end{aligned} \tag{2}$$

The first two terms in (2) are the approximation errors. The third term $\eta\|A^1 - A^2\|^2$ penalizes the difference between A^1 and A^2, and the last term $\lambda(\|A^1\|^2 + \|A^2\|^2)$ contains the shrinkage regularizers that help avoid overfitting. The parameters η and λ balance the fitting accuracy and the generalization performance.

Please refer to [13] for more details of the CSR method.

3.2. Deep Coupled Spectral Regression

Although CSR delivers a more discriminative subspace for heterogeneous face data than CCA and its variants do [13] [14], its performance heavily relies on the design of hand-crafted features for facial representation. To solve this problem, we propose to incorporate deep features, *i.e.* CNNs, into CSR, which are reputed to be more powerful to automatically learn non-linear representations from raw data.

We denote the features of different modalities generated by two different CNNs as $F^1 = \Phi_1(X^1)$ and $F^2 = \Phi_2(X^2)$ respectively. The objective function of CSR can then be rewritten as

$$\begin{aligned} L_{CSR} = & \frac{1}{n} \sum_i^n (\|Y_i^1 - {f_i^1}^T A^1\|^2 + \|Y_i^2 - {f_i^2}^T A^2\|^2) \\ & + \eta \|A^1 - A^2\|^2 + \lambda(\|A^1\|^2 + \|A^2\|^2). \end{aligned} \tag{3}$$

where n is the batch size, $f_i^1 \in F^1$ and $f_i^2 \in F^2$.

The gradient based optimization algorithm is widely used in training the neural network. The keypoint in optimizing the objective function is to compute its gradient

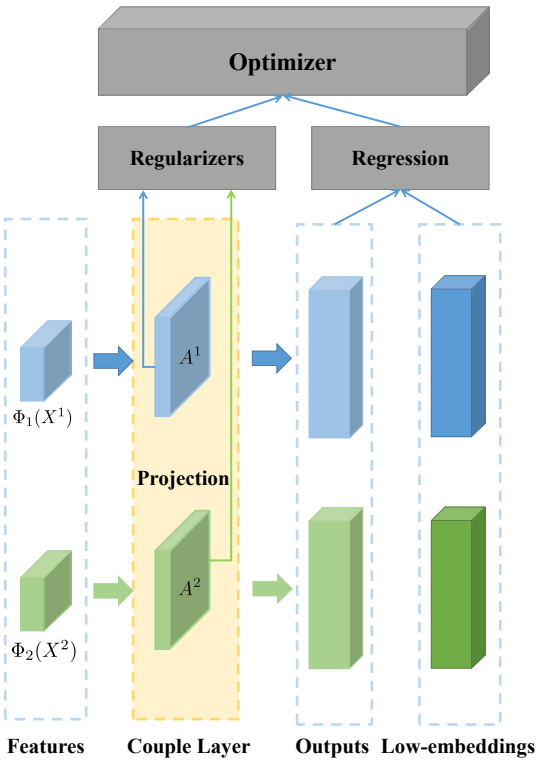


Figure 2. Architecture of the couple layer. It takes two input layers and generates two output layers. Meanwhile it delivers the shrinkage regularizer to the optimizer.

with respect to the projection matrices A^1 and A^2. We use L_{CSR} to represent the objective function, and its gradient with respect to A^1 can be derived as

$$\begin{aligned}\frac{\partial L_{CSR}}{\partial A^1} &= \frac{1}{n}\sum_i^n\left(\frac{\partial\|Y_i^1 - {f_i^1}^T A^1\|^2}{\partial A^1} + \frac{\partial\|Y_i^2 - {f_i^2}^T A^2\|^2}{\partial A^1}\right) \\ &+ \eta\frac{\partial\|A^1 - A^2\|^2}{\partial A^1} + \lambda\frac{\partial(\|A^1\|^2 + \|A^2\|^2)}{\partial A^1} \\ &= \frac{2}{n}\sum_i^n f_i^1({f_i^1}^T A^1 - Y_i^1) \\ &+ 2\eta(A^1 - A^2) + 2\lambda A^1.\end{aligned} \tag{4}$$

Similarly, the gradient of the objective function with respect to A^2 can be written as

$$\begin{aligned}\frac{\partial L_{CSR}}{\partial A^2} &= \frac{2}{n}\sum_i^n f_i^2({f_i^2}^T A^2 - Y_i^2) \\ &+ 2\eta(A^2 - A^1) + 2\lambda A^2.\end{aligned} \tag{5}$$

3.3. Couple Layer

Different from traditional neural networks, our DCSR method has the objective function that contains constraints on the projection matrices, namely the Shrinkage regularizer that avoids overfitting and the regularizer which penalizes the difference between these matrices.

For joint optimization in feature extraction and mapping learning, we build a simple yet effective layer structure, called the couple layer. The overall structure is shown in Figure 2. The couple layer contains the two projection matrices A^1 and A^2, and has two input layers. The output of the couple layer is the same as that of the traditional fully connected layer, which employs linear transformation to process data. Moreover, we deliver the aforementioned two regularizers to the optimizer in the couple layer.

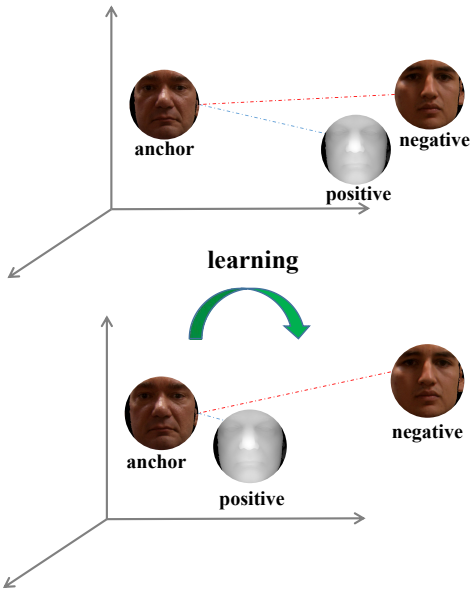


Figure 3. The cross-modality triplet loss encourages the faces of the same identity to be projected closer to each other, while enlarges the margin between the ones of different identities.

3.4. Cross-modality Triplet Loss

Inspired by [1, 18], we design a cross-modality triplet loss to learn more discriminative facial representations in 2D-3D HFR, which preserves the intra-personal similarity and enlarges the inter-personal margin in different modalities.

Given $g_i \in F^T A = \{{F^1}^T A^1, {F^2}^T A^2\}$ that represents the low-embeddings of the two different modalities after projection, we randomly sample a set of cross-modality triplet tuples (g_i^a, g_i^p, g_i^n), where the anchor sample g_i^a and the positive sample g_i^p are features of the same identity (they can come from different modalities), while the negative one g_i^n is from a different identity.

The cross-modality triplet loss is then defined as:

$$L_{CMTL} = \frac{1}{m}\sum_i^m[\|g_i^a - g_i^p\|^2 - \|g_i^a - g_i^n\|^2 + \alpha]_+ \tag{6}$$

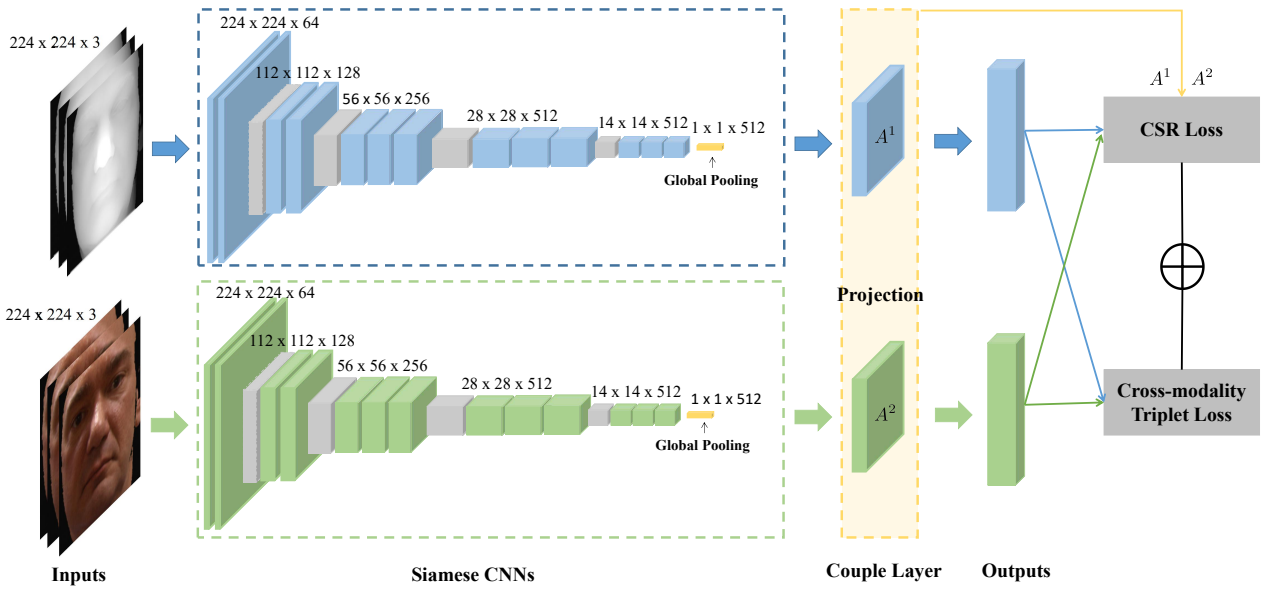


Figure 4. Overview of the proposed DCSR based 2D-3D HFR method.

where m is the number of the cross-modality triplet tuples, α is the margin between samples of different identities, and $[\cdot]_+$ refers to $max(\cdot, 0)$.

The selection of the cross-modality triplet tuples is based on the following constraints:

$$\begin{cases} \|g_i^a - g_i^p\|^2 < \|g_i^a - g_i^n\|^2 \\ \|g_i^a - g_i^n\|^2 < \|g_i^a - g_i^p\|^2 + \alpha \\ g_i^a, g_i^p, g_i^n \in F^T A = \{{F^1}^T A^1, {F^2}^T A^2\} \end{cases} \quad (7)$$

Under this constraint, all faces of the same identity (no matter what modal they come from) are encouraged to be projected closer to each other in the embedding space, while the margin between the samples of different identities is enlarged.

Figure 3 illustrates the concept of this cross-modality triplet loss.

3.5. Network Architecture

Based on (3) and (6), we combine the CSR term and the cross-modality triplet loss term as our final objective function:

$$L_{DCSR} = L_{CSR} + L_{CMTL} \quad (8)$$

The whole structure of DCSR is shown in Figure 4. The network framework is similar to the CNN configuration provided in [23]. The ConvNet is composed of five sequences of convolutional layers: two 64-dims, two 128-dims, three 256-dims, three 512-dims, and three 512-dims. We use ReLU as the activation function. The size of all the receptive fields is set as 3×3 and the stride in all the layers is 1. A max pooling layer is inserted between each pair of sequences. In the last convolutional layer, instead of the fully connected layer, we use the global pooling layer [17], which significantly reduces the feature dimensionality. The couple layer follows the global pooling layer in the end.

4. Experiments

To evaluate the proposed method for 2D-3D HFR, we conduct extensive experiments on the FRGC v2.0 database [20]. To further validate its generalization ability, we carry out additional experiments on the CASIA NIR-VIS 2.0 database [16].

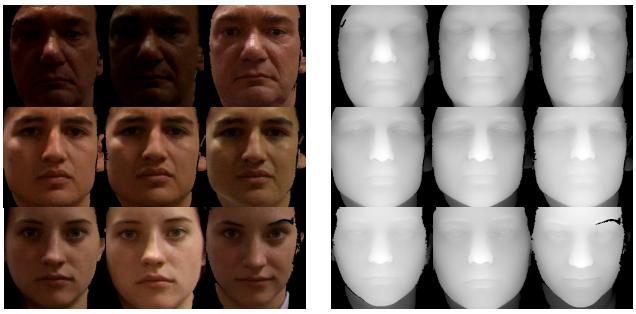
The databases, parameter settings, protocols, and results are described in the subsequent subsections.

4.1. Database

FRGC v2.0 is one of the most famous and comprehensive databases to evaluate 2D-2D FR, 3D-3D FR, and 2D-3D HFR approaches. It consists of 4,007 textured 3D face models of 466 subjects, possessing large changes in facial expressions, illumination conditions and moderate variations in head poses. The facial texture and range images separately extracted from 3D models are resized to 224×224 pixels as input. Figure 5 shows some examples of texture and range images from this database.

4.2. Protocols and Settings

To analyze the effectiveness and highlight the advantage of our method, we design two experimental protocols,



(a) Texture images (b) Range images

Figure 5. Examples of facial texture and range images on the FRGC v2.0 database: (a) texture images and (b) range images.

Table 1. Comparison of 2D-3D HFR methods in the person independent experiment on FRGC v2.0.

Method	Rank-1 RR %	VR@FAR=0.001 %
PCA + CCA [5]	30.22	15.91
PCA + CSR [13]	47.40	23.86
OGMs + CCA [7]	58.01	41.68
OGMs + CSR [7][13]	71.26	53.76
Deep Feature + CCA [5]	63.31	32.24
Deep Feature + CSR [13]	75.18	49.84
Deep CCA [24]	69.78	49.21
DCSR	**95.97**	**87.70**

for the person independent (PI) and person dependent (PD) problems.

Protocol I: For PI, no identities are shared in the training and test sets. This experiment aims to simulate the open-set situation, which requires high quality of the common subspace produced as well as a strong generalization ability. We take the last 366 subjects according to the order of subject ID that have 2,964 images for each view as training data to fine-tune CNN. The range images of the first 3D models of the remaining 100 subjects compose the gallery set, and the texture images of the rest 3D models of the 100 subjects (943 images) are used as probes.

Protocol II: For PD, identities in the test set also appear in the training set. This experiment is to simulate the close-set situation, where high recognition accuracy is requested for enrolled persons. Therefore, the gallery set can be included as part of the training set. Besides the 2,964 images from the last 366 subjects, we add the first texture and range images from the gallery set into the training set, which totally contains 3,064 images for each modality. The test set is the same as that in Protocol I.

In both the protocols, different kinds of features are employed to represent 2D and 3D face images. In all the experiments, the structure of CNN is described in Sec.3.5, and the batch size and learning rate are set as 32 and 1e-5 respectively. The Labeled Faces in the Wild (LFW) dataset [12] is used to pre-train the CNNs, and the training data on FRGC v2.0 are used for fine-tuning.

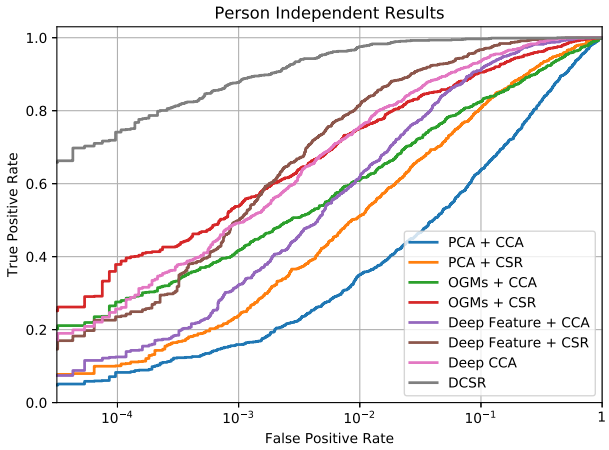


Figure 6. ROC curves of different methods in the person independent experiment

In the classification phase, the cosine similarity is employed to describe the relationship among different features in the common sub-space. At last, we use Nearest Neighbor (NN) as the classifier to make decision. The experimental results are reported with two standard indices, Rank-1 Recognition Rate (RR) and Verification Rate (VR) with the False Acceptance Rate (FAR) at 0.001. The Receiver Operating Characteristic (ROC) curves of all the experiments are provided to evaluate the performance of different methods.

The proposed DCSR is compared with pixel feature based CCA [5] and CSR [13], OGMs feature based CCA [7] and CSR [13], Deep feature (extracted from the VGG-Face model [23]) based CCA [5] and CSR [13], and the latest Deep CCA model [24] under the same settings. It should be noted that the best hyper-parameters of η and λ are set in CSR according to [13]. In DCSR, $\{\eta, \lambda, \alpha\}$ are set as $\{1e\text{-}5, 1e\text{-}6, 1.0\}$ respectively.

4.3. Personal Independent Results

Table 1 illustrates the results under Protocol I (PI). It can be seen from the experimental scores that the traditional methods which separate feature extraction and cross-modality mapping are generally unable to report satisfactory performance in the open-set 2D-3D FR problem. The best accuracy among those methods only reaches 75.18% (through a simple combination of deep CNN features and CSR) for identification. Deep CCA improves the performance of direct integration of CNN and CCA, from 63.31% to 69.78%, which shows the superiority of the end-to-end trainable deep model. Compared with them, our DCSR achieves the performance of 95.97%, with an approximate 20% promotion. Regarding the verification task, DCSR

Table 2. Comparison of 2D-3D HFR methods in the person dependent experiment on FRGC v2.0.

Method	Rank-1 RR %	VR@FAR=0.001 %
PCA + CCA [5]	44.64	27.89
PCA + CSR [13]	66.38	48.37
OGMs + CCA [7]	66.91	54.61
OGMs + CSR [7][13]	80.06	66.70
Deep Feature + CCA [5]	90.99	74.13
Deep Feature + CSR [13]	96.08	92.05
Deep CCA [24]	97.56	97.99
DCSR	**99.26**	**98.83**

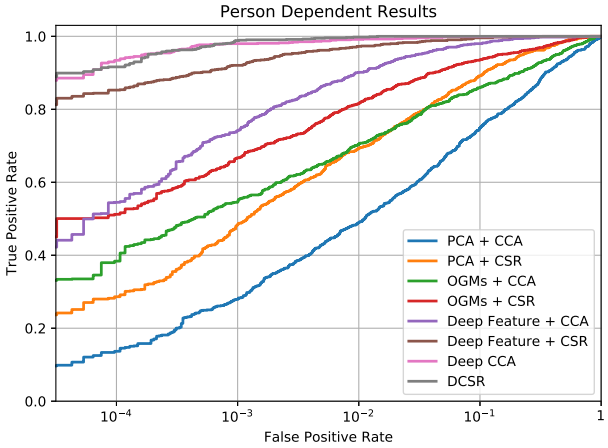


Figure 7. ROC curves of different methods in the person dependent experiment

shows a VR of 87.70% @FAR=0.001, which is the best score among all the counterparts as well.

It is worth noting that all the CSR based methods perform better than CCA based ones using the same features, which indicates that with the help of supervised information, the shared subspaces learned from CSR are more discriminative than the ones of the unsupervised methods. Additionally, thanks to the cross-modality triplet loss and the couple layer with which the joint optimization is conducted, the advantages of CNN and CSR are sufficiently exploited, greatly outperforming their simple combination.

As shown in Figure 6, the proposed DCSR method achieves significantly better performance than the counterparts in the person independent experiment.

4.4. Personal Dependent Results

The results of different methods in protocol II (PD) are displayed in Table 2. CSR based approaches learn more discriminative subspaces for different facial representations than CCA based approaches do. Table 2 shows that the methods with deep learning frameworks achieve better scores in both identification and verification compared to the traditional methods using hand-crafted features. The best performance of shallow features is only 80.06% for Rank-1 RR and 66.70% for VR@FAR=0.001, while the accuracies of deep feature based methods are dramatically improved. The DCSR model reports the best performance, 99.26% for Rank-1 RR in identification and 98.83% for VR@FAR=0.001 in verification.

Figure 7 shows the ROC curves in the person dependent experiment. Due to the additional information provided by sample labels, the DCSR model learns more discriminative representations for both the two different modalities, leading to better precisions.

Table 3. Comparison with different NIR-VIS HFR methods on CASIA NIR-VIS 2.0

Method	Rank-1 %	VR@FAR=0.001 %
TRIVET [18]	95.7 ± 0.5	91.0 ± 1.3
IDR [3]	97.3 ± 0.4	95.7 ± 0.7
W-CNN [4]	$\mathbf{98.7 \pm 0.3}$	$\mathbf{98.4 \pm 0.4}$
DCSR	98.4 ± 0.3	97.8 ± 0.4

4.5. Experiments for NIR-VIS HFR

To further validate the generalization ability of our proposed method, additional experiments are conducted in the NIR-VIS HFR task on the CASIA NIR-VIS 2.0 database [16].

CASIA NIR-VIS 2.0 is the largest and most challenging NIR-VIS HFR database due to the large variations in lighting, expression, pose and distance. The database contains 725 subjects captured with 1-22 VIS and 5-50 NIR face images per subject. In the training phase, there are about 2,500 VIS and 6,100 NIR images from 357 subjects. In the testing phase, the gallery set is constructed from 358 subjects and each identity only has a single VIS image. The probe set contains over 6,000 NIR images from the same 358 subjects. The subjects in the training and testing are different and the two sets are disjointed. Thus, it is an open-set recognition problem.

We compare the performance of our DCSR method with that of the other state-of-the-art NIR-VIS HFR methods, such as TRIVET[18], IDR [3] and W-CNN [4]. For fair comparison, the light CNN model [25] is used as the backbone network, pre-trained on the MS-Celeb-1M dataset [2]. 10-fold cross validation experiments are conducted according to the standard protocol described in [16].

Table 3 shows the results of different methods on CASIA NIR-VIS 2.0. The performance of the proposed DCSR method is $98.4 \pm 0.3\%$ for Rank-1 RR and $97.8 \pm 0.4\%$ for VR@FAR=0.001. They are comparable to the best scores reported in [4] for NIR-VIS HFR, which demonstrate the good generalization ability of our method for other HFR problems.

5. Conclusion

In this paper, we propose a novel end-to-end deep learning based approach for 2D-3D HFR, namely Deep Coupled Spectral Regression (DCSR), which incorporates the advantages of both CNN and CSR. For joint optimization in feature extraction and mapping learning, we build a simple yet effective couple layer. A cross-modality triplet loss is designed to further enhance the discriminative power of facial features in different modalities. We validate our method on FRGC v2.0 in two different scenarios of identification and verification. The experiments with the person independent and person dependent protocols, are carried out. The proposed DCSR model reaches state-of-the-art performance compared to the counterparts in 2D-3D HFR. The results on CASIA NIR-VIS 2.0 also show that the proposed method can be well generalized to other HFR problems.

Acknowledgment

This work is funded by the National Natural Science Foundation of China under Grant 61673033.

References

[1] S. Florian, K. Dmitry, and P. James. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[2] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102, 2016.

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

[4] R. He, X. Wu, Z. Sun, and T. Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE TPAMI*, 2018.

[5] D. Huang, M. Ardabilian, Y. Wang, and L. Chen. Asymmetric 3D-2D face recognition based on lbp facial representation and canonical correlation analysis. In *ICIP*, pages 3325–3328, 2009.

[6] D. Huang, M. Ardabilian, Y. Wang, and L. Chen. 3-d face recognition using elbp-based facial description and local feature hybrid matching. *IEEE TIFS*, 7(5):1551–1565, 2012.

[7] D. Huang, M. Ardabilian, Y. Wang, and L. Chen. Oriented gradient maps based automatic asymmetric 3D-2D face recognition. In *ICB*, pages 125–131, 2012.

[8] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. Automatic asymmetric 3D-2D face recognition. In *ICPR*, pages 1225–1228, 2010.

[9] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. Local binary patterns and its application to facial image analysis: A survey. *IEEE TSMC-C*, 41(6):765–781, 2011.

[10] D. Huang, W. Ben Soltana, M. Ardabilian, Y. Wang, and L. Chen. Textured 3D face recognition using biological vision-based facial representation and optimized weighted sum fusion. In *CVPR Workshops*, pages 1–8, 2011.

[11] D. Huang, Y. Wang, and Y. Wang. A robust method for near infrared face recognition based on extended local binary pattern. In *ISVC*, pages 437–446, 2007.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[13] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *CVPR*, pages 1123–1128, 2009.

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

[15] H. Li, D. Huang, J. Morvan, Y. Wang, and L. Chen. Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors. *IJCV*, 113(2):128–142, 2015.

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

[17] M. Lin, Q. Chen, and S. Yan. Network in network. *ICLR*, 2013.

[18] X. Liu, L. Song, X. Wu, and T. Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *ICB*, pages 1–8, 2016.

[19] S. Ouyang, T. Hospedales, Y. Song, X. Li, C. C. Loy, and X. Wang. A survey on heterogeneous face recognition. *Image and Vision Computing*, 56:28–48, 2016.

[20] P. J. Phillips, P. J. Flynn, K. W. Bowyer T. Scruggs, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR*, pages 947–954, 2005.

[21] A. Rama, F. Tarres, D. Onofrio, and S. Tubaro. Mixed 2D-3D information for pose estimation and face recognition. In *ICASSP*, 2006.

[22] D. Riccio and J. L. Dugelay. Asymmetric 3D-2D processing: a novel approach for face recognition. In *ICIAP*, pages 986–993, 2005.

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[24] S. Wang, D. Huang, Y. Wang, and Y. Tang. 2D-3D heterogeneous face recognition based on deep canonical correlation analysis. In *CCBR*, pages 77–85, 2017.

[25] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE TIFS*, 13(11):2884–2896, 2018.

[26] W. Yang, D. Yi, Z. Lei, J. Sang, and S. Z. Li. 2D-3D face matching using CCA. In *FG*, pages 1–6, 2008.

[27] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3dface: A large-scale database of low-cost kinect 3d faces. In *ICB*, pages 1–8, 2016.

[28] W. Zhang, Z. Shu, D. Samaras, and L. Chen. Improving heterogeneous face recognition with conditional adversarial networks. *arXiv preprint arXiv:1709.02848*, 2017.

[29] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520, 2011.