Transferring Deep Representation for NIR-VIS Heterogeneous Face Recognition

Xiaoxiang Liu[1,2], Lingxiao Song[1], Xiang Wu[1], Tieniu Tan[1]
[1]Center for Research on Intelligent Perception and Computing,
Institute of Automation, Chinese Academy of Sciences, Beijing, P. R. China, 100190
[2]School of Automation,
Harbin University of Science and Technology, Harbin, P. R. China, 150080
xiaoxiang.liu@cripac.ia.ac.cn, {lingxiao.song, tnt}@nlpr.ia.ac.cn, alfredxiangwu@gmail.com

Abstract

*One task of heterogeneous face recognition is to match a near infrared (NIR) face image to a visible light (VIS) image. In practice, there are often a few pairwise NIR-VIS face images but it is easy to collect lots of VIS face images. Therefore, how to use these unpaired VIS images to improve the NIR-VIS recognition accuracy is an ongoing issue. This paper presents a deep **T**ransfe**R** N**I**R-**V**IS heterogeneous fac**E** recognition ne**T**work (TRIVET) for NIR-VIS face recognition. First, to utilize large numbers of unpaired VIS face images, we employ the deep convolutional neural network (CNN) with ordinal measures to learn discriminative models. The ordinal activation function (Max-Feature-Map) is used to select discriminative features and make the models robust and lighten. Second, we transfer these models to NIR-VIS domain by fine-tuning with two types of NIR-VIS triplet loss. The triplet loss not only reduces intra-class NIR-VIS variations but also augments the number of positive training sample pairs. It makes fine-tuning deep models on a small dataset possible. The proposed method achieves state-of-the-art recognition performance on the most challenging CASIA NIR-VIS 2.0 Face Database. It achieves a new record on rank-1 accuracy of **95.74**% and verification rate of **91.03**% at FAR=0.001. It cuts the error rate in comparison with the best accuracy [27] by **69**%.*

1. Introduction

Illumination variations often significantly degenerate the recognition performance of a visible light (VIS) face recognition system. Near infrared (NIR) imaging technique provides an effective solution to acquire images robust to lighting variations with a low cost. Owing to the superior characteristic of NIR images, face or other biometrics [8, 25, 24] can serve better in security surveillance and authentication applications. The motivation behind NIR-VIS heterogeneous face recognition is to bridge the gap between the face images generated under lights of two different wavelength ranges.

Biological vision can amazingly create some hazy feelings and connotations like romanticism art so that humans can easily distinguish identities no matter how abstract the subjects presented in front of our eyes are. However, it is hard for computers to effectively learn the highly non-linear relationship and other coupled variations between two different modalities. Much attention has been poured into the research of face recognition so far, both with traditional methods and deep learning methods. However, heterogeneous face recognition especially NIR-VIS face recognition has not been well explored despite its extensive potential applications.

Traditional methods on heterogeneous face recognition mainly focus on three strategies to alleviate the cross-modal gap [13]: designing invariant features for different modalities, transforming one face modality to the other, and projecting both image modalities onto a common subspace. Modal-invariant features SIFT or LBP are extracted in [4, 5]. Synthesis based approaches are employed in the research of [20, 10]. Tang *et al*. [20] propose an Eigen-transformation method while Liu *et al*. [10] reconstruct image patches using LLE. Approaches of [22] and [14] project cross-domain images to a common subspace by employing LDA and TCA (transfer component analysis), respectively. Recently, Felix *et al*. [3] propose a ℓ_0-Dictionary based approach to reconstruct face images on the basis of images in the other domain, which achieves the highest verification rate (85.80%) on the CASIA NIR-VIS 2.0 Face Database[7] up to now.

With the development of deep learning method, many vision related problems enter into a new era. Some attempts have been made with respect to heterogeneous face recognition. J. Ngiam *et al*. [12] propose a Bimodal Deep AE method based on denoising antoencoder. To exert the potential effects of all layers, Srivastava *et al*. [18] suggest a

978-1-5090-1869-7/16/$31.00 ©2016 IEEE

multi-modal DBM approach. The state-of-the-art rank-1 accuracy (86.16%) is achieved by [27], which resorts to RBM combined with removed PCA features. Although these unsupervised approaches often perform well on small-scale NIR-VIS datasets, the matching accuracy of NIR-VIS is still far below than those of the VIS face recognition methods. Convolutional neural networks (CNNs), as a branch of deep learning methods, take advantage of volumes of labeled data and have the ability to extract high-level features with its hierarchical structure. However, there are two ongoing issues for employing deep CNNs to NIR-VIS research: (1) How to utilize millions of unpaired VIS face images to improve the accuracy of NIR-VIS recognition? (2) How to apply convolutional neural networks on a small number of paired NIR-VIS dataset?

To address these two issues, this paper studies a deep transfer learning method for NIR-VIS heterogeneous face recognition. We utilize the convolutional neural network with ordinal measures (o-CNN) [23]. The ordinal activation function (Max-Feature-Map, MFM) in [23] potentially makes the CNN robust to heterogeneous variations. The large-scale CASIA WebFace Database [26] is used to provide pre-training on o-CNN for extracting general face features. The learned filters offer a prior knowledge for NIR-VIS domain data. As for the problem of limited NIR-VIS images, we take the triplet loss [17] into consideration. Specifically, we come up with two types of NIR-VIS triplet loss to reduce intra-class NIR-VIS variations, and meanwhile augment positive training sample pairs. The triplet loss can impose a strong constraint on lots of distinct combinations between NIR-VIS images, which makes fine-tuning deep models on a small dataset possible. As a result, we can fine-tune on the pre-trained network with the triplet loss to make the model adapt to task-specific data. Experimental results on the most challenging CASIA NIR-VIS 2.0 Face Database show that the proposed approach performs better than state-of-the-art NIR-VIS face recognition methods. Experimental results also demonstrate that o-CNN learns unified features for NIR-VIS domain and provides a potential candidate for small-scale heterogeneous problems.

To sum up, this paper makes three main contributions:

(1) We first propose a deep transfer CNN learning method for NIR-VIS heterogeneous face recognition problem, which can efficiently utilize large-scale unpaired VIS images to deal with the recognition problem on small-scale paired NIR-VIS heterogeneous images.

(2) Different from the fact that fine-tuning deep models in a new domain requires a large amount of labeled data [21], our method employs NIR-VIS triplets to enlarge the positive sample space, which provides a new strategy for the small sample problem in NIR-VIS.

(3) Experimental results on the most challenging CASIA NIR-VIS 2.0 Face Database demonstrate that the proposed method improves state-of-the-art rank-1 accuracy from 86.16% to **95.74**% and VR (@FAR=0.001) from 85.80% to **91.03**%.

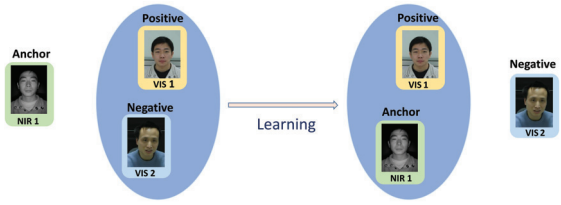

Figure 1. NIR-VIS triplet formation. After learning, the distance between the samples from the same ID is minimized while the difference across domains tends to be not the leading factor. [Best viewed in color]

In the rest of our paper, we will elaborate the proposed framework in Section 2, describe the details and results of our experiments in Section 3 and discuss in Section 4. Section 5 concludes this paper.

2. Deep Transfer Architecture for NIR-VIS

Biometric applications based on CNNs explode fast in recent years. With the hierarchical structure, CNNs have the ability to learn the deep representation for input images. Due to limited NIR-VIS image sources, heterogeneous face recognition problem has not been dabbled in the supervised method based on CNNs. In light of the perfect performance of CNNs on VIS face recognition, fine-grained object recognition, object classification, etc., we make an attempt to enlarge our training data by constructing lots of triplets. Furthermore, a prior knowledge is given to the network by pre-training on the large-scale CASIA WebFace Dataset. To alleviate the effect of degradation brought by the shift across domains, we then fine-tune the learned model on the specific NIR-VIS data to learn a domain-invariant deep representation [21, 11]. In the subsections, we will introduce the triplet loss first, and then give a detailed explanation to our TRIVET architecture.

2.1. Triplets Formation

A triplet is composed of an anchor, a positive exemplar and a negative exemplar. The triplet loss can be formulated as Eq. 1.

$$loss = \sum_{i}^{N} \left[||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha \right]_+ \tag{1}$$

To give a constraint on the distance amongst x_i^a (anchor), x_i^p (positive) and x_i^n (negative), one minimizes the sum of the loss. The loss imposes a constraint on the distance between intra-subject and inter-subject. With this type of loss, some kind of nonlinear projection can be learned through SGD

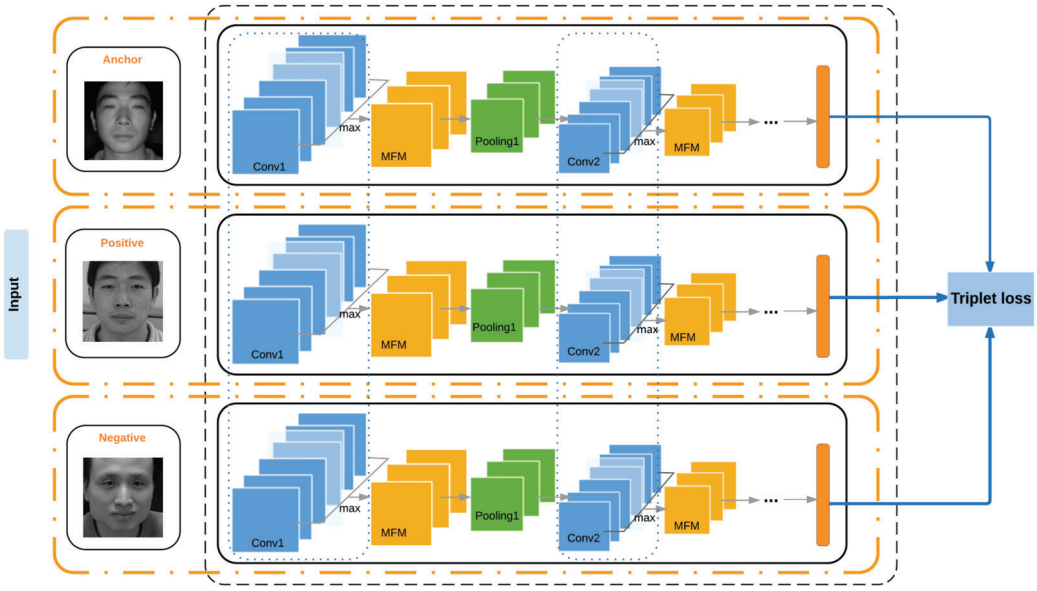

Figure 2. The proposed CNN framework. The inputs of CNN are the prepared triplets, and the three channels share the same parameters. After feature extraction in the final fully connected layers, the high-level features of the three layers are input to the triplet loss layer which bridges the gap of NIR and VIS domains.

(stochastic gradient descent). Consequently, intra-class distance decreases to a minimum with a margin, while inter-class distance reaches a maximum [17].

It is worth mentioning that, the random selection of triplets for single-modal images is not applicable for cross-domain data, because such triplet formation may not reduce the gap between different modalities. To learn the discriminative representation for domain-invariant data, we suggest a novel selection strategy for cross-domain samples and provide two types of NIR-VIS triplet loss. Firstly, we set a near infrared image as an anchor, a visible image of the same ID as a positive example and another visible image with a different ID as a negative example (see Fig. 1). Then, to make the relative relationship of NIR and VIS balanced, we do the similar procedure to generate triplets whose anchor is a VIS image, but positive and negative samples are from NIR images. With the constraint of these two kinds of cross-domain triplet loss, the network focuses more on the individual distinction so that cross-domain variance is weakened or even eliminated. Thus in the hyperspace, no matter which modality the face image belongs to, face features are enforced to be closer if they are from the same identity and far away from each other if they belong to different identities.

2.2. Hard NIR-VIS Triplets Selection

The selection of hard triplets is suggested in [17] to avoid slow convergence. Hard samples are the images with impertinent similarity scores computed by the network. When they are picked out for re-training, the network will pay more attention to face features that are hard to distinguish for the model and adjust the model towards the direction adapting more to hard triplets.

For cross-domain NIR-VIS data, to generate hard samples, we input all training images (both NIR and VIS images) to the test network to compute the similarity matrix. Take a NIR anchor as an example (the anchor can also be a VIS image), its similarity scores with all VIS face images are computed. The images with lower scores but from the same identity with the anchor are regarded as hard positive samples, while the images with higher scores but from different IDs are marked as hard negative samples. The hard triplets are composed of a probe NIR photo, a hard positive VIS image and a hard negative VIS image. Fig. 3 illustrates the strategy for the selection of hard triplets. Several triplets are generated for each anchor to increase the discriminative ability of the model.

2.3. CNN with Ordinal Measures

General CNNs employ sigmoid or ReLU as the activation function to acquire nonlinear transformations. Inspired by maxout networks [2], the CNN with ordinal measures (o-CNN) [23] comes up with an ordinal activation function named MFM (Max-Feature-Map) that directly compares ordinal relationships between different neurons. MFM defines a sparse connection between two convolutional layers by an ordinal measure, i.e., extracting the maximum of candidate

nodes in two feature maps, which takes the following form,

$$f_{ij}^{k} = \max(C_{ij}^{k}, C_{ij}^{k+n}), k \in \{1, ..., n\} \quad (2)$$

where C denotes the convolutional layer with $2n$ channels, i and j represent the spatial location of the filters. Such a strategy can not only make the network lightened but also select remarkable features. Since the relationship between NIR and VIS is highly non-linear, MFM fits a lot. Compared with ReLU and its variants, the representation generated by MFM is compact and the gradients of MFM are sparse. The ordinal structure of o-CNN decreases the number of parameters and makes it potentially useful for small-scale NIR-VIS training data. Another merit of o-CNN is its fast running speed.

The lightened CNN [23] consists of 4 convolutional layers, each with a MFM activation and a max pooling layer. After that, two fully connected layers are cascaded to generate deep representative features. The input size of the network is 144×144 and then images are randomly cropped to 128×128. The first convolutional layer has 96 response maps with filter size of 9×9. The second convolutional layer creates 192 outputs with filter size of 5×5. In the third convolutional layer, the size of the filter is also 5×5 and it generates 256 feature maps as outputs. 384 feature maps are generated by the last convolutional layer with filter size of 4×4. Response maps of each convolutional layer are divided averagely into two parts as inputs of MFM. The size of 4 max pooling layers is identically 2×2.

To provide entries to image triplets for training, we design a three-channel CNN architecture based on [23]. To reduce the computational complexity and the size of parameters, we only employ one fully connected layer to generate features of 512 dimensions. Fig. 2 shows the whole framework. The three channels share the same weights at each layer. After the final fully connected layer (with MFM transformation), the outputs of these three channels are input to a triplet loss layer. With the constraint of the triplet loss, discriminative features can be learned to differentiate different identities no matter which modality they belong to, NIR or VIS. The shared weights of three channels indicate the relationship between an input face image (NIR or VIS) and its corresponding high-level features, and hence give a unified representation for both NIR and VIS images. As a result, our model is also suited for heterogeneous recognition where VIS and NIR images are used as probes and gallery respectively. To be consistent with the standard protocols of the CASIA NIR-VIS 2.0 Database, we only conduct experiments using NIR images as probes.

2.4. Deep Transfer Learning

To address the problem of small-scale data size of paired NIR-VIS face images, we assume that the near infrared images and visible photos possess the equal discriminability

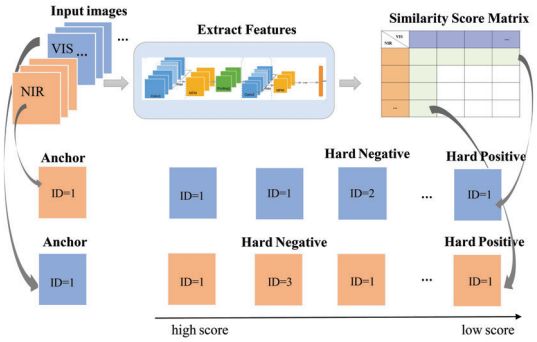

Figure 3. The selection of hard triplets. All NIR and VIS images are input to one trained channel to extract high-level features, then their similarity score matix is computed. For each anchor, images with improper scores are selected as hard positive and negative samples. The task is many-to-many so there are more than one image for each ID.

for the same person. Thus we can take advantage of numbers of visible face images to pre-train our model. Here we choose the CASIA WebFace Dataset [26] which has 494,414 images from 10,575 identities. Even so, we cannot use the learned matrix to attain the domain-invariant representation for NIR-VIS data directly, because the distribution shift between near infrared data and visible data will bring about the problem of performance degradation. Suppose the network has learned the significant features of a face representation after pre-training, what we need to do is to adapt the general face recognition network to our task-specific NIR-VIS face images.

By fine-tuning on the new data, our network learns not only the essential characteristics identifying who a person is, but also the nonessential characteristics ignoring particular modality the input image pertains to. This process shares much similarity with reasoning and generalization abilities of humans. When a baby see a plant with an elongated trunk, supporting branches and leaves for the first time, he will be told that it is a tree. Later on, he may point at a willow or a pine tree and shout: "Tree, its a tree!". What the baby focuses on is not the detailed shape of the tree (which can be described as some kind of heterogeneity), but the prominent features that differentiate trees from grasses or flowers. Deep transfer learning is also a process making efforts to capture the essential characteristics of an identity while ignoring the specific modality.

3. Experiments

3.1. Dataset and Protocols

In order to evaluate the performance of the proposed method, experiments are conducted on the most challenging CASIA NIR-VIS 2.0 Face Database [7]. The database is

widely used in NIR-VIS heterogeneous face evaluations because it is the largest publicly available dataset across NIR and VIS spectrums up to now. This database collects a total number of 725 subjects, each with 1-22 VIS and 5-50 NIR images. There are not one-to-one correlations between the images in NIR and VIS domains. The images are randomly gathered. It is extremely challenging due to various variations of the same identity, including lighting, expression, pose, and distance. Variations are also generated by wearing glasses or not.

Two views of evaluations are provided in the CASIA 2.0 Face Database. View 1 is used for super-parameters adjustment and View 2 acts as normal training sets and test sets. To ensure the randomness, View 2 divides the whole dataset to 10 sub-experiments, each contains a collection of training, test, gallery and probe lists. To have a fair comparison with other's results, we adopt standard protocols as well. Training on the dataset is many-to-many (i.e., images from NIR and VIS are randomly combined). When the test phase is executed, one VIS image from each person is used to compose the gallery while any NIR images in the database can be used as probes (many-to-one). Nearly equal numbers of identities are included in the training and test sets, and are kept disjoint from each other.

3.2. Experimental Preparations

To get more representative face images and utilize the deep representation generated by the pre-trained network, we crop and align original images in the CASIA NIR-VIS 2.0 Face Database anew. We use a face detector [19] to detect the face, and then locate three landmarks: centers of two eye pupils and the middle of the mouth. The similarity transformation is executed firstly between two pupil centers and then between the mid-point of two eyes and the center of the mouth. After that, face images are cropped and resized to 144×144 pixels. For network pre-training, the images in the CASIA WebFace Database are pre-processed in the similar way, thus we can transfer the pre-trained model to new NIR-VIS data seamlessly.

For each sub-experiment, we generate new training triplets with the principle of covering more possible combinations randomly and uniformly. The anchor can be a NIR or VIS image while the positive and negative samples are in the opposite modality with the same ID or not. The total number of triplets is nearly 200,000 for each sub-experiment (we traverse all images and multiply different times for NIR and VIS to 100,000 respectively).

3.3. Experimental Results

To illustrate our hypothesis that enlarged triplet samples can help small-scale data training for deep CNNs, we utilize the prepared triplets without any further processing (just with random matching and no hard samples are selected) and directly input them to the pre-trained network. After the first round fine-tuning, the designed system is capable of embedding the deep face representation for specific NIR and VIS images. Hence we test the model using one of the three channels to get deep features for both VIS gallery images and NIR probe images. Then cosine distances between features are calculated as the similarity scores to judge whether the gallery and probe images from different domains are of the same ID.

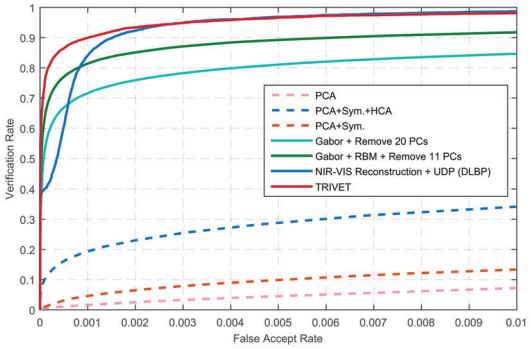

Figure 4. The ROC curves of different NIR-VIS face recognition methods.

For each sub-experiment, we compute the rank-1 accuracy and VR@FAR=0.1% (to be specific, here VR denotes TPR) respectively. Then the average values and the standard deviations of rank-1 recognition accuracy and verification rate for the ten sub-experiments are reported. The results in Table 1 show that our method with random selection of triplets surpasses all other methods with rank-1 accuracy of 91.48% and VR of 86.80%. These results also verify that the triplet loss is a reasonable loss function for small-scale data.

Since the fine-tuned model with random selection of triplets behaves well on the CASIA NIR-VIS 2.0 database, we conduct the hard sample selection to further validate the performance of our method. After being picked out, hard samples are input to the network to refine the results. The process is iterated for several times until the performance no longer increases. At last, the proposed method outperforms the state-of-the-art methods with the rank-1 accuracy of 95.74% which improves the second highest result by 9.58%. The best VR of 91.03% is also achieved by our approach which surpasses the second highest result by 5.23%. Higher rank-1 accuracy and VR indicate a better recognition and discrimination ability and good results also illustrate the effectiveness of our unique TRIVET framework.

Fig. 4 plots the ROC curves of different NIR-VIS face recognition methods. We observe that when FAR is larger than 0.001, the methods can be ordered in descending verification rates as TRIVET, DLBP, Gabor + RBM + PCA, Gabor + PCA, PCA + Sym. + HCA, PCA + Sym., and PCA.

Table 1. Experimental results on the CASIA 2.0 NIR-VIS Face Database in terms of rank-1 accuracy and VR@FAR=0.1%.

	Rank-1 Accuracy	VR@FAR=0.1%
PCA+Sym+HCA [7]	$23.70 \pm 1.89\%$	19.27%
Cognitec [1]	$58.56 \pm 1.19\%$	-
DSIFT+LDA [1]	$73.28 \pm 1.10\%$	-
Gabor + RBM + Remove 11 PCs [27]	$\mathbf{86.16} \pm 0.98\%$	$81.29 \pm 1.82\%$
NIR-VIS Reconstruction + UDP (DLBP) [3]	$78.46 \pm 1.67\%$	**85.80%**
OURS(without fine-tuning)	$79.01 \pm 1.62\%$	$69.48 \pm 1.90\%$
OURS(fine-tuning+triplet loss)	$91.48 \pm 1.19\%$	$86.80 \pm 1.80\%$
OURS(fine-tuning+triplet loss+hard sample selection)	$\mathbf{95.74} \pm 0.52\%$	$\mathbf{91.03} \pm 1.26\%$

When FAR is smaller than 0.001, the proposed TRIVET significantly outperforms its competitors. It achieves the state-of-the-art of 91.03%, and surpasses the second highest VR (DLBP) by 5.23 percent at FAR=0.1%. It is also interesting to observe that when FAR is smaller than 0.001, DLBP performs no better than Gabor + RBM + PCA and Gabor + PCA methods. This may be due to that Gabor features often provide discriminative informations.

Table 1 further tabulates rank-1 accuracy and VR@FAR=0.1% of different NIR-VIS face recognition methods. Conclusion can be made that TRIVET performs the best compared with other methods. It reaches the highest rank-1 accuracy of 95.74%, which has a promotion up to 9.58 percent compared with the second highest value of Gabor + RBM + PCA. The significant improvements obtained by TRIVET may be because TRIVET can efficiently utilize both large-scale unpaired VIS face images and small-scale paired NIR-VIS face images simultaneously. We also contrast the results without fine-tuning. As expected, the deep transfer process and the unique triplet loss are two indispensable strategies for promoting the performance of NIR-VIS heterogeneous face recognition. This can be explained by that discriminative domain-invariant features are learned after fine-tuning with the constraint of the triplet loss.

3.4. Experimental Analysis

We visualize feature maps in Conv1 (the first convolutional layer), Conv2, Conv3 and Conv4 respectively (Fig. 5). The inputs of feature maps in each layer are one NIR face image and one VIS face image from the same identity. For the purpose of saving space, only 4×4 feature maps corresponding to 16 filters are showed for each convolutional layer. From each modality, we can get the following observations: clear face outlines can be perceived from feature maps in Conv1, while there are some phantoms in Conv2. Due to the effects of subsampling caused by pooling operation, feature maps in Conv3 are much more blurry than Conv2, but we can still recognize with some efforts that they are faces because of their rough sketch and the legible shapes and layouts of eyes, noses and mouths. In Conv4,

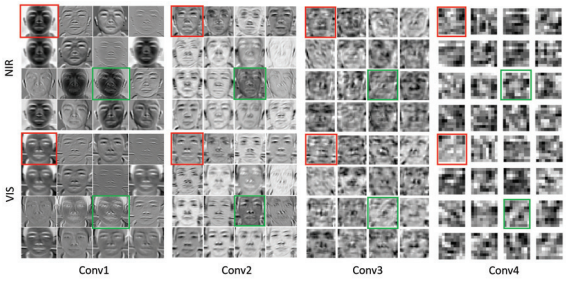

Figure 5. Feature maps in 4 convolutional layers. In the first row, feature maps are from NIR face images. In the second row, feature maps are from VIS face images. Comparing the corresponding NIR and VIS feature maps in red boxes and green boxes, we can see that the differences across domains are decreased as the neural network become deeper. [Best viewed in color]

the visualized feature maps are 9×9 for each filter and it is hard to observe any information of a face. Such perceptions illustrate that the hierarchical structure of CNNs can learn features automatically from input images and as the number of layers increases, more high-level and abstract the features are.

High-level features can help to minimize the differences across domains, which can be demonstrated from the comparison of the first row and the second row for each convolutional layer in Fig. 5. From Conv1 to Conv4, the differences between the near infrared face image and its counterpart, the visible face image, become more and more imperceptible. For example, in position (1, 1) (bounded by red boxes) and (3, 3) (bounded by green boxes) of NIR and VIS feature maps, conspicuous differences exist between two modalities in Conv1, but with deeper network, more similarities are shared in Conv2 and Conv3, and no obvious differences can be perceived in Conv4. The fact illustrates that the transition of face features is from general to specific for our task [28]. Such observations also demonstrate that our TRIVET can learn a unified deep representation for face images in both near infrared domain and visible light domain, so it is effective and brilliant for the NIR-VIS heterogeneous face matching problem.

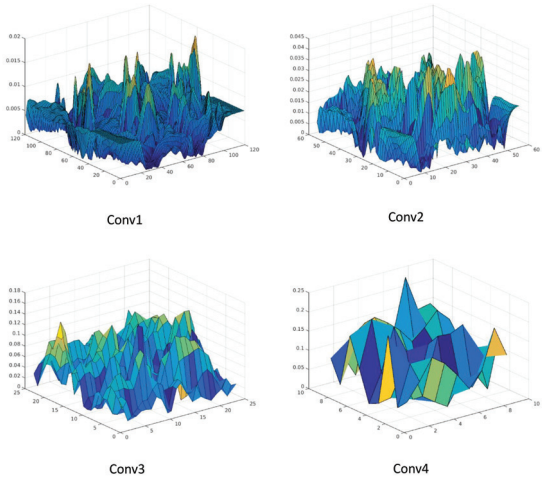

Figure 6. The visualization of absolute difference between feature maps of an input NIR face image generated before and after fine-tuning.

Fig. 6 further plots the absolute difference value between the feature maps generated with pre-trained model and fine-tuned model in 4 convolutional layers. For each convolutional layer, the feature maps come from one filter with one NIR face image as an input, and are of size (120×120), (56×56), (24×24), (9×9) respectively. X-coordinate and Y-coordinate denote the spatial position of feature maps and the absolute differences between pixels respectively, which decide the value of Z-coordinate. The scale range of Z-coordinate tends to be bigger as convolutional layers become higher. This indicates that the fine-tuning process with the triplet loss has more influence over high-level features in the network. It can be explained that the shallow features of the face images in visible database and specific NIR-VIS database are similar because the filters tend to learn general and common representations such like edges and corners in lower layers. However, with the constraint of the triplet loss, after transferring, the parameters of the model are tuned to adapt to the new training data. As a result, the high-level features in deeper layers become specified and task-relevant. This observation suggests that the fine-tuning and the cross-domain triplet loss facilitate the process of domain-invariant features extraction.

Some examples of misclassified NIR-VIS pairs are showed in Figure 7. We conjecture that the false negative samples might mainly caused by large pose variations. As for the false positive samples, it is very hard to distinguish the pairs even for humans.

4. Discussion

For NIR-VIS face image matching task, the label for a pair of input images is either 1 or 0 and it tends to be 0

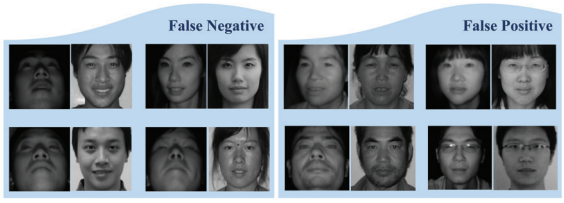

Figure 7. The samples of misclassifed NIR-VIS pairs.

in most cases. If a learning method treats all input images as different identities, it will minimize the whole error rate rather than give special attention to positive classes. This may lead to a low verification rate but a relatively high accuracy [29]. Triplet loss has shown its merits in our method and some related work [15]. The main reason may lie in a stronger constraint in ternary relations than in binary relations. A potential trend may be that tetrad or quintuple samples would result in better performance.

A hot topic rises gradually on how an algorithm generalizes well when it encounters new inputs with only one or several labeled examples. Humans can learn a concept with one initial positive instance and enhance the recognition ability by learning. It is a further attempt for computer vision to imitate human vision mechanism and learning process. To address this problem, several methods have been proposed both in Bayesian learning method [6] and deep learning method [9, 16] with one or only several labeled inputs, but with complex learning mechanisms. Our proposed architecture provides a small trial towards this trend with small-scale data for training and relatively simple and easy learning process. More efforts need to be poured into this theme and we can expect that future computer vision can handle one-input learning efficiently.

5. Conclusion

In this paper, we propose a deep transfer NIR-VIS heterogeneous face recognition network named TRIVET for matching NIR and VIS face images. The proposed architecture is an unified framework that integrates the deep representation transferring and the triplet loss to get consolidated feature representations for face images in two modalities. Experimental results on the CASIA 2.0 NIR-VIS Face Database show that the proposed method achieves significant improvements and is effective for extracting common high-level features of face images from different domains. Our proposed method can alleviate the over-fitting problem for CNNs on small-scale datasets. Although experiments are implemented on NIR-VIS data in this paper, the presented approach can also be applied to other cross-domain heterogeneous recognition problems.

6. Aknowledgements

This work is supported by the Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS) (Grant No. 2015190), the National Natural Science Foundation of China (Grant No. 61473289) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02070000).

References

[1] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa. On effectiveness of histogram of oriented gradient features for visible to near infrared face matching. In *Pattern Recognition, IEEE International Conference on*, pages 1788–1793, 2014.

[2] I. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1319–1327, 2013.

[3] F. Juefei-Xu, D. K. Pal, and M. Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *Computer Vision and Pattern Recognition Workshops, IEEE International Conference on*, pages 141–150, 2015.

[4] G. H. K. and S. T. Inter-modality face sketch recognition. In *Multimedia and Expo, IEEE International Conference on*, pages 224–229, 2012.

[5] B. Klare and A. K. Jain. Sketch-to-photo matching: a feature-based approach. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7667, page 1, 2010.

[6] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[7] S. Z. Li, D. Yi, Z. Lei, and S. Liao. The casia nir-vis 2.0 face database. In *Computer Vision and Pattern Recognition Workshops, IEEE International Conference on*, pages 348–353, 2013.

[8] S. Z. Li, L. Zhang, S. Liao, X. Zhu, R. Chu, M. Ao, and R. He. A near-infrared image based face recognition system. In *Automatic Face and Gesture Recognition, IEEE International Conference on*, pages 455–460, 2006.

[9] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Computational baby learning. In *arXiv preprint arXiv:1411.2861*, 2014.

[10] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 1, pages 1005–1010, 2005.

[11] M. Long and J. Wang. Learning transferable features with deep adaptation networks. In *arXiv preprint arXiv:1502.02791*, 2015.

[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Machine Learning, International Conference on*, pages 689–696, 2011.

[13] S. Ouyang, T. Hospedales, Y. Song, and X. Li. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. In *arXiv preprint arXiv:1409.5114*, 2014.

[14] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *Neural Networks, IEEE Transactions on*, 22(2):199–210, 2011.

[15] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *arXiv preprint arXiv:1506.02640*, 2015.

[17] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clusteringmann machines. In *Computer Vision and Pattern Recognition, IEEE International Conference on*, pages 815–823, 2015.

[18] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Neural Information Processing Systems (NIPS)*, pages 2222–2230, 2012.

[19] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition, IEEE International Conference on*, pages 3476–3483, 2013.

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

[21] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Computer Vision, IEEE International Conference on*, pages 4068–4076, 2015.

[22] R. Wang, J. Yang, D. Yi, and S. Z. Li. An analysis-by-synthesis method for heterogeneous face biometrics. In *Biometrics, International Conference on*, pages 319–326, 2009.

[23] X. Wu, R. He, and Z. Sun. A lightened cnn for deep face representation. In *arXiv preprint arXiv:1511.02683v1*, 2015.

[24] L. Xiao, Z. Sun, R. He, and T. Tan. Coupled feature selection for cross-sensor iris recognition. In *Biometrics: Theory, Applications and Systems, IEEE International Conference on*, pages 1–6, 2013.

[25] L. Xiao, Z. Sun, R. He, and T. Tan. Margin based feature selection for cross-sensor iris recognition via linear programming. In *Pattern Recognition, IEEE Asian Conference on*, pages 246–250, 2013.

[26] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. In *arXiv preprint arXiv:1411.7923*, 2014.

[27] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Shared representation learning for heterogeneous face recognition. In *Automatic Face and Gesture Recognition, IEEE International Conference and Workshops on*, volume 1, pages 1–7, 2015.

[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Neural Information Processing Systems (NIPS)*, pages 3320–3328, 2014.

[29] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji. Deep model based transfer and multi-task learning for biological image analysis. In *Knowledge Discovery and Data Mining, ACM SIGKDD Conference on*, pages 1475–1484, 2015.