IMAGE-IMAGE TRANSLATION TO ENHANCE NEAR INFRARED FACE RECOGNITION

Fangyu Wu[1,2]*, Weihang You*[2]*, Jeremy S. Smith*[1]*, Wenjin Lu*[2]*, Bailing Zhang*[3]

[1]Department of Electrical Engineering and Electronics, University of Liverpool
[2]Department of Computer Science and Software Engineering
Xi'an Jiaotong-Liverpool University
[3]The Institute of Advanced Artificial Intelligence in Nanjing

ABSTRACT

With the rapid development of facial recognition, the research field of near infrared (NIR) face recognition, which is less sensitive to illumination levels, has attracted increased attention. Unfortunately, directly applying the face recognition model trained using visual light (VIS) data to NIR face data does not produce a satisfactory performance. This is due to the domain bias between the NIR image and the VIS image. To this end, we created the Outdoor NIR-VIS Face (ONVF) database and Indoor NIR Face (INF) database to increase the number of near infrared facial images. In this paper, we propose an efficient NIR face recognition method, which consists of face detection and alignment, NIR-VIS image translation and face embedding. The NIR-VIS image conversion model is capable of transforming near-infrared facial images into their corresponding VIS images whilst maintaining sufficient identity information to enable existing VIS facial recognition models to perform recognition. Extensive experiments using the INF Dataset and the CSIST Database have demonstrated that the proposed method yields a consistent and competitive performance for near infrared face recognition.

Index Terms— Near infrared face recognition, Image-image translation, Face embedding

1. INTRODUCTION

Facial recognition has become one of the most active research areas in the field of computer vision due to its potential value for many applications such as security systems and surveillance. Despite significant progress in this domain, illumination has been regraded as one of the most significant impact factors in face recognition [1]. In this paper, we focus on near infrared (NIR) face recognition, which has the features of being insensitive to illumination changes and can perform well even in near darkness [2]. Many algorithms have been proposed to recognize faces in NIR images in recent years [3], [4]. A common drawback of all these methods is that they exploited hand-crafted features without applying a deep, global representation of the facial images, which has been shown to produce superior results for face recognition.

Our work is motivated by two recent developments. Firstly, the existing visible light domain (VLD) face recognition systems have achieved impressive performance [5], [6], owning to the development of deep networks and large face datasets [7], [8]. For example, in [5], 200 million images captured from 8 million subjects were used to train a deep network which achieve the best performance on a standard unconstrained face recognition benchmark called Labeled Faces in the Wild (LFW) [7]. With these significant advances, existing VLD-based face recognition systems should be extended to other research areas which are less studied, such as near-infrared imaging (low-light). Unfortunately, due to the relatively small amount of training data available and the domain bias between near-infrared and visible light, the same success in VLD is not easily replicated in the near-infrared domain. This observation inspired us to resort to synthesize visible light face images from NIR inputs, which solves the illumination change problem and, at the same time, is able to work with pre-existing face recognition systems.

Secondly, with the development of the Generative Adversarial Network (GAN) method, the community has made significant progress in solving image-image translation problems [9], [10]. Recently, SPGAN [11] introduced the Siamese network based on CycleGAN framework in [10] to learn the image translation between two different domains and preserve the identity information of the person. However, these GAN based methods require sufficient input-output image pairs for training, which is not available for the near infrared domain with limited samples. To address this problem, it is important to introduce datasets that include sufficient near infrared and visual light image pairs.

Considering the above two issues, this paper makes two contributions. The first contribution is the creation of two new NIR datasets, named the "Outdoor NIR-VIS Face (ONVF) database" and "Indoor NIR Face (INF) database". The ONVF dataset contains 30,000 image pairs of 1,000 identities collected by a visible light camera and a near-infrared camera. The INF dataset consists of 470 near-infrared images which belong to 94 people captured by a near-infrared camera. The the details of the databases are descrbed in Section 3.1.

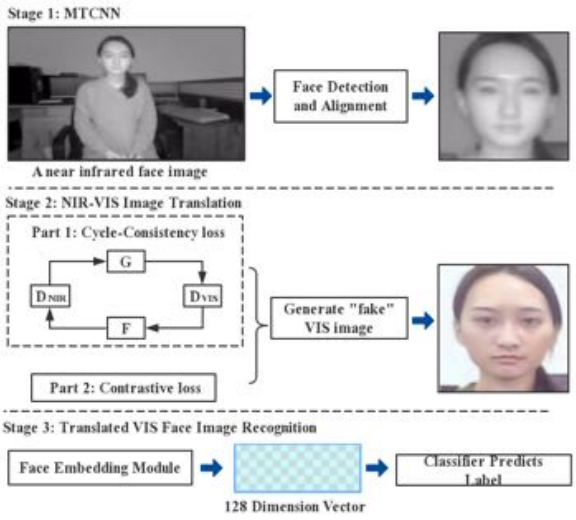

Fig. 1. The workflow of the proposed approach. There are three stages: (1) Face detection and alignment using MTCNN [12]; (2) NIR-VIS Image translation by combining the cycle-consistency loss and contrastive loss; (3) Translated VIS face image recognition which obtains a 128 dimension feature vector for each translated VIS face image in the test set. Finally, the predict label will be output by the classifier.

As a second contribution, we propose a novel near infrared facial recognition method. To start, a Multi-task Cascaded Convolutional Network (MTCNN) [12] is applied to achieve face detection and alignment, which is useful for handling background and occlusion variations in facial images. Then, the ONVF dataset is used to train a NIR-VIS image translation model that translates the near infrared face image to a visible light face image. After the translation, the generated VLD face is fed into an existing pre-trained VLD deep neural network face recognition model [5]. The intention is that better recognition results can be obtained without retraining or changing the VLD model. The framework of our proposed method is illustrated in Figure 1. The experimental results on the INF and CSIST [13] database confirm that the proposed method achieves a favorable performance compared with published state-of-the-art methods [21], [22].

2. PROPOSED METHOD

In this section, the proposed method is described, which consists of three steps: 1) Face detection and alignment, 2) NIR-VIS image translation, 3) Face embedding and classification.

2.1. Face Detection and Alignment

Zhang et al. [12] proposed the MTCNN framework that exploits the inherent correlation between detection and alignment to improve their performance. In our work, we apply the MTCNN network to implement a face detection model on the ONVF dataset and CSIST [13]. It essentially consists of three stages: (1) We exploit a list of candidate windows which are generated by the proposed network (P-Net) to classify the face and non-face and estimate the bounding box regression vector as the face position. (2) A large number of wrong candidates will be rejected by feeding all the candidates to a Refining Network (R-Net). (3) Another CNN, called O-Net, outputs the five facial landmarks.

2.2. NIR-VIS Image Translation

The goal of NIR-VIS image translation is to train a generator G that can transform the near infrared image X into its corresponding visible image Y, where the visible image Y contains sufficient identity information for the facial recognition task. To this end, we utilised image-image translation methods which aim at learning a mapping function between the two domains. Conditional GAN [15] is a representative method by using paired training data to produce impressive transition results. However, it is difficult to obtain sufficient paired training data in the real world. In [10], this framework has been extended to unsupervised image-to-image translation, meaning there is no requirement for image pairs. In this work, we applied the CycleGAN framework [10] to transform a NIR image to a VIS image. Two generator-discriminator pairs are introduced, G, D_{NIR} and F, D_{VIS}, which map a sample from the NIR domain to the VIS domain and produce a sample that is indistinguishable from those in the VIS domain. For generator G and its associated discriminator D_{VIS}, we express the adversarial loss as

$$\begin{aligned}\mathcal{L}_{VIS_{adv}}(G, D_{VIS}, P_x, P_y) =& E_{y\sim p_y}[(D_{VIS}(y)-1)^2]+ \\ & E_{x\sim p_x}[(D_{VIS}(G(x))^2],\end{aligned} \tag{1}$$

where p_x and p_y denote the sample distributions in the NIR and VIS domains, respectively. For generator F and its associated discriminator D_{VIS} , the adversarial loss is

$$\begin{aligned}\mathcal{L}_{NIS_{adv}}(F, D_A, P_y, P_x) =& E_{x\sim p_x}[(D_A(x)-1)^2]+ \\ & E_{y\sim p_y}[(D_A(F(y))^2],\end{aligned} \tag{2}$$

Due to the lack of paired training data, there exists multiple alternative mapping functions. To restore the original image after a cycle of translation and reverse translation, we then introduced a cycle-consistency loss [10] as:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) =& E_{x\sim p_x}[||F(G(x))-x||_1]+ \\ & E_{y\sim p_y}[||G(F(y))-y||_1],\end{aligned} \tag{3}$$

Similarity preservation is an important principle to exploit synthesised images generation from some GAN-based image-image translation schemes. In our work, additional constraints have been set on the mapping function to meet this special requirement for face image generation. Specifically,

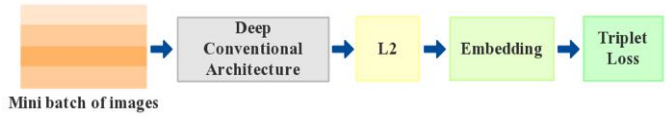

Fig. 2. The Face Embedding Module.

inspired by the success in [11], we add the contrastive loss [16] in the cycle-consistency loss function to learn a latent space that constrains the learning of the mapping function.

$$\mathcal{L}_{con}(l, i_1, i_2) = (1-l)\{max(0, m-d)\}^2 + ld^2, \quad (4)$$

where i_1 and i_2 are a pair of input vectors, which are selected in an unsupervised manner. d denotes the cosine distance between the normalized embedding of two input vectors, and l represents the binary label of the pair. If i_1 and i_2 are a positive image pair, l equals one. On the contrary, if i_1 and i_2 are a negative image pair, l equals zero. Suppose two samples denoted as x_{NIR} and x_{VIS} come from the NIR domain and VIS domain, respectively. We define two positive pairs: 1) x_{NIR} and $G(x_{NIR})$, 2) x_{VIS} and $F(x_{VIS})$. The positive image pairs contain the same person, the only difference is that they have different styles (NIR or VIS). In the learning procedure, we encourage the whole network to pull these two images close. There are also two types of negative training pairs designed for generators G and F: 1) $G(x_{NIR})$ and x_{VIS}, 2) $F(x_{VIS})$ and x_{NIR}.

The separability in the embedding space has been represented as $m \in [0, 2]$. When m equals zero, there is no back-propagation for the negative training pair. If m is larger than zero, the system will consider the loss of both the positive and negative sample pairs. A larger m means that the loss of negative training samples have a higher weight for the back propagation. Taken together, the final NIR-VIS translation objective can be written as in equation (5) by considering Eqs (1), (2), (3), and (4):

$$\mathcal{L}_{sum} = \mathcal{L}_{B_{adv}} + \mathcal{L}_{A_{adv}} + \mathcal{L}_{cyc} + \mathcal{L}_{con} \quad (5)$$

2.3. Face Embedding Module

A high-performance facial embedding module is critical to the entire facial recognition system. In this paper, our face embedding module is based on FaceNet [5] which is a deep measurement learning network that uses two different convolutional neural network structures, [17] and [18]. We use the Inception-ResNet-v1 model, which achieves similar precision but with fewer parameters and lower computational complexity. First, the face-embedded module has been considered as a black box (Figure 2), and the Inception-ResNet-v1 model is the most important part of this end-to-end system. There is a batch input layer and deep CNN (Inception-ResNet-v1) in our network, which is then followed by L2 standardized for face embedding. The training network is then trained through the triples loss [17]. The basic idea is that the distance between the vectors of facial images from the same person is very small.

3. EXPERIMENTS

3.1. Datasets Introduction

ONVF database. The database was collected by our research team in the summer session of 2018 and the procedure lasted for several days. We captured the facial images from 1,000 subjects without constraints on the illumination and pose, each subject providing about 30 NIR and 30 VIS face images. There are various variations in the sample images such as pose, expression and focus, etc. During the process of collecting the database, we use JAI cameras with 1/2.7inch HM2131 image sensor which is sensitive to the NIR band. The active light source was in the NIR spectrum between 780nm - 1,100nm and it was mounted on the camera. We applied the full ONVF database to train a NIR-VIS image translation model.

INF database. The INF database consists of 94 students from the University, including 57 male and 37 females. During the recording, a subject was asked to sit in front of the camera, and their normal frontal face images were collected. The camera-face distance was between 80-120 cm, a convenient range for the user. There are 5 NIR face images per subject at a resolution of 640×480 pixels. In our experiments, we randomly selected 235 images for training and 235 images for testing.

CSIST database. There are two image sets in the CSIST database [13]: Lab1 and Lab2. Lab1 consists of 500 NIR images and 500 visible images captured from 50 subjects. In Lab2, 1000 NIR images and 1000 visible images are collected from 50 subjects under different illumination conditions. The image sizes for the Lab1 and Lab2 databases is 100×80 pixels. We randomly selected 50% of the database images for training, and 50% for testing.

3.2. Implementation Details

Firstly, we implemented the facial detection and alignment using MTCNN [12], the facial images generated by the face pre-processing are 640×640 pixels. Then, we use the large ONVF database to train our NIR-VIS image translation model in Tensorflow [19]. During training, we set the initial learning rate and training epoch as 0.0002 and 7, respectively. During the testing procedure, we employed the trained NIR-VIS image translation model to translate the NIR face images in the INF database and CSIST database to visible facial images. Examples of images translated by NIR-VIS image translation are shown in Fig. 3. Also, to carry out comparative studies, we also trained the CycleGAN using the same settings.

Table 1. NIR face image recognition accuracy (%) on the INF database

Method	INF database
Plain Near Infrared	79.8
CycleGAN	34.9
MTCNN+CycleGAN	97.5
Proposed method	**99.8**

Table 2. NIR face image recognition accuracy (%) on the CSIST database.

Method	Lab 1	Lab 2
SMRSN [21]	-	86.4
Score-level fusion [22]	-	88.63
Plain Near Infrared (ours)	65.2	77.2
CycleGAN (ours)	45.4	38.6
MTCNN+CycleGAN (ours)	94.2	86.3
Proposed method (ours)	**99.6**	**90.7**

For the Face Embedding Module, we extracted an image-level CNN based on the Inception-ResNet-v1 network [20] which was trained on a subset of the MSCeleb-1M database [8] and validated on the LFW database [7]. The model's architecture follows the Inception-ResNet-v1 network [20]. The input image size of the Inception-ResNet- v1 model is 160×160 pixels.

3.3. Results and Discussion

To help analyze our model and show the benefit of each module, we designed three baselines as follows:

Plain Near Infrared. No transformation was applied on this baseline. This baseline will indicate the effect of the domain gap between NIR and VIS on the face recognition models trained solely using VIS images.

CycleGAN. We first train the CycleGAN (without contrastive loss) using the ONSF Database, and the generated visual face images are fed into the face embedding model.

MTCNN+CycleGAN. Before training the CycleGAN, we added the face detection and alignment step by using MTCNN.

The results on the INF dataset can be seen in Table 1. As can be seen from these results the proposed method outperforms all the baselines. Also, it is noteworthy to mention that the MTCNN leads to a 62.6% improvement which indicates that the face detection and alignment is able to further improve the system performance. The improvement brought by the contrastive loss is also validated on this dataset, with a 2.3% improvement over the second baseline.

We then recorded the recognition accuracy of our methods on the CSIST database as shown in Table 2. The MTCNN+CycleGAN method has an Average Precision (AP) value of 94.2% and 86.3% for Lab1 and Lab2, respectively,

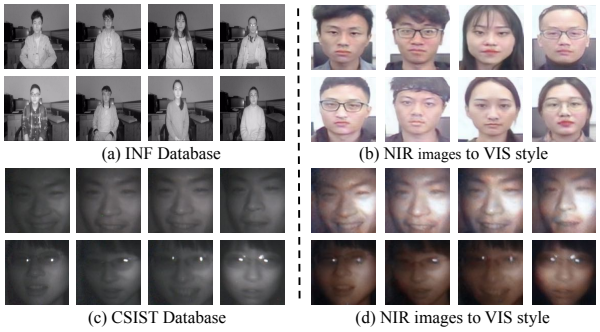

Fig. 3. Example images translated from NIR to VIS.

which is higher than the Plain Near Infrared and CycleGAN performance. This highlights the importance of reducing the domain bias that exists between the NIR and VIS images. With the help of contrastive loss, we preserve the identity information during the image translation process leading to a 5.4% and 4.4% improvement over the second baseline for Lab1 and Lab2 versions respectively.

The following observations can be made: (1) For the CSIST database, our method outperformed most of the previous methods which are only based on the near infrared domain. (2) The NIR-VIS image translation model is more effective at training a generator that can preserve the personal identity in the generated images. (3) The proposed model shows good potential to achieve better results. Future work can be undertaken by fine-tuning the face embedding model on a specific dataset.

4. CONCLUSION

In this paper, we introduced the large ONSF database which includes various changes in pose, expressions and focus. We also created another INF database in the laboratory environment to test the performance of near infrared facial recognition. We propose a novel near infrared facial recognition method in an end-to-end deep architecture which includes face detection and alignment, NIR-VIS image translation and a face embedding module. This is the first time that it has been proposed to apply the image-image translation method to enhance the performance of a near-infrared facial image recognition. This is achieved by synthesizing a virtual sample from an input near infrared face image. Using this approach, we reduce the intra-personal difference caused by the completely different illumination. Therefore, we can achieve much better recognition results by applying the existing pre-trained VLD deep neural network face recognition model. The proposed method was tested on the INF database and the CSIST dataset, with promising results.

5. REFERENCES

[1] Anil K Jain and Stan Z Li, *Handbook of face recognition*, Springer, 2011.

[2] Christopher Reale, Nasser M Nasrabadi, Heesung Kwon, and Rama Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *CVPRW*, 2016, pp. 54–62.

[3] Baochang Zhang, Lei Zhang, David Zhang, and Linlin Shen, "Directional binary code with application to polyu near-infrared face database," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2337–2344, 2010.

[4] S Farokhi, UU Sheikh, J Flusser, SM Shamsuddin, and H Hashemi, "Evaluating feature extractors and dimension reduction methods for near infrared face recognition systems," *Jurnal Teknologi*, vol. 70, pp. 23–33, 2014.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[6] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[7] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*. Springer, 2016, pp. 87–102.

[9] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 700–708.

[10] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2242–2251.

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification," in *CVPR*, 2018, vol. 1, p. 6.

[12] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[13] Yong Xu, Aini Zhong, Jian Yang, and David Zhang, "Bimodal biometrics based on a representation and recognition approach," *Optical Engineering*, vol. 50, no. 3, pp. 183–183, 2011.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*. IEEE, 2017, pp. 5967–5976.

[16] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006, pp. 1735–1742.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[18] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[19] Martn Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard, "Tensorflow: a system for large-scale machine learning," 2016.

[20] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *AAAI*, 2017, vol. 4, p. 12.

[21] Mohamed Anouar Borgi, Demetrio Labate, Maher El Arbi, and Chokri Ben Amar, "Sparse multi-regularized shearlet-network using convex relaxation for face recognition," in *International Conference on Pattern Recognition*, 2014.

[22] Guodong Liu, Shuai Zhang, and Zhihua Xie, "A novel infrared and visible face fusion recognition method based on non-subsampled contourlet transform," in *CISP-BMEI*. IEEE, 2017, pp. 1–6.