FINE-TUNING APPROACH TO NIR FACE RECOGNITION

Jeyeon Kim[1] *Hoon Jo*[1] *Moonsoo Ra*[1] *Whoi-Yul Kim*[2]

[1] Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea
2 Department of Electronic Engineering, Hanyang University, Seoul, Korea
E-mail: jykim@vision.hanyang.ac.kr

ABSTRACT

Despite extensive researches for face recognition (FR), it is still difficult to apply deep CNN models to NIR FR due to a lack of training data. In this study, we propose a fine-tuning approach to allow deep CNN models to be applied to NIR FR with small training datasets. In the proposed approach, parameters of deep CNN models for RGB FR are utilized as initial parameters to train deep CNN models for NIR FR. The proposed approach has two main advantages: 1) High NIR FR performances can be achieved with very small public training datasets. 2) We can easily secure good generalization for NIR FR in various environments. Our fine-tuning approach achieved a validation rate of 99.70% with the PolyU-NIRFD database. In addition, we constructed private face databases with Intel® RealSense™ SR300. On the VF_NIR database, which is one of the private databases, we achieved a validation rate of 94.47%.

Index Terms— Face verification, face identification, biometrics, deep learning, transfer learning

1. INTRODUCTION

Face recognition (FR) is one of the most actively researched biometrics in the field of computer vision. The performance of RGB FR has been improved by more than 99% by developing deep convolutional neural network (CNN) models [1-6]. However, RGB face images have a disadvantage in that the intensity of the face part of the image is largely dependent on the lighting environment. Especially, as shown in Fig. 1, the intensity of RGB face images changes drastically in smartphone unlocking scenarios, such as on a dark street or indoors.

To overcome the above-mentioned disadvantage of RGB images, several studies on near-infrared (NIR) FR have been conducted, because NIR face images are less affected by the illumination environment due to the active NIR lights of NIR sensors. Recently, NIR FR based on deep CNN models has been researched to improve the performance of face recognizers to as good as those for RGB FR [7-13]. In existing NIR FR approaches [14, 15], deep CNN models with simple structures have been introduced, and these approaches achieve good performances (95% or above) [14, 15]. However, since the performance evaluation of the existing approaches was conducted on NIR face databases, which are constructed in limited environments, the same high performance cannot be expected in real-world FR scenarios, which involve various environments. Therefore, to improve the generality of NIR FR, we must apply complex deep CNN architectures [1, 4, 5, 8] to NIR FR; such architectures have been shown to be powerful for RGB FR in various environments. However, when complex deep CNN architectures are directly applied to NIR FR, we cannot expect good performance because the architectures are insufficiently trained by public NIR face databases. The size of current NIR face databases is only one tenth of the well-known CASIA WebFace database [18] for RGB FR.

Fig. 1. RGB face images in several lighting conditions. The images were acquired assuming real smartphone unlocking scenarios. In (a) and (b), the face images were captured in an outdoor environment with bright and weak lighting. (c) shows the image captured in an indoor environment with weak lighting. (d) was captured in the similar environment as (c) but with the shadow on the face. (e) was captured in a room with extremely weak lighting.

To solve this issue, we have used a fine-tuning approach that utilizes pre-trained information from a deep CNN model for RGB FR to improve the performance of NIR FR. More specifically, the proposed fine-tuning approach utilizes parameters of the pre-trained RGB model as the initial parameters of the NIR deep CNN model. The NIR deep CNN model means the deep CNN model for NIR FR, and the pre-trained RGB model means the deep CNN model which is trained by RGB face images before the NIR deep CNN model is trained. In addition, we have shown how the pre-trained RGB model can be effectively used to solve NIR FR problems. In the experimental results section, we discuss the performance evaluation which we conducted by applying off-the-shelf deep CNN architectures to the fine-tuning approach using NIR face images, and we show which deep CNN architecture has the best performance in NIR FR among the architectures. In addition, the performance of the proposed fine-tuning approach is compared with existing methods such as NIR FR [14, 15] and RGB FR. Finally, we show that the proposed approach has better generalization ability for various environments in a real-world FR scenario than existing NIR methods [14, 15] and RGB FR.

2. PROPOSED FINE-TUNING APPROACH

In this study, the NIR FR method is the same as FaceNet [10]. An image pair, which includes two NIR face images, is used as an input to the NIR face recognizer. If the input image pair includes

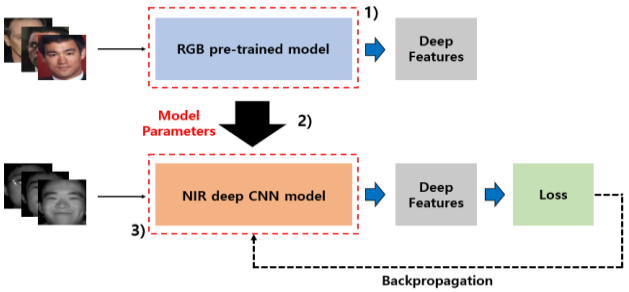

Fig. 2. The NIR deep CNN model training method based on the proposed fine-tuning approach

two faces of the same person, it is a positive pair; otherwise, it is a negative pair. Deep features of the input pair are extracted from the NIR deep CNN model, and NIR FR is conducted using the Eclidean distance of the deep features. If the value of Eclidean distance is smaller than a threshold, the input pair is recognized as a same person. Otherwise, the input pair is recognized as a different person.

2.1 Training NIR Deep CNN Model

The proposed fine-tuning approach is adopted to train the NIR deep CNN model using only tens of thousands of NIR face images. Fine-tuning is a transfer learning method [16, 24, 25]. Transfer learning indicates that the information acquired to solve one problem is utilized to solve another. In the proposed fine-tuning approach, the information of the pre-trained RGB model is used to train the NIR deep CNN model. Fig. 2 shows the proposed fine-tuning approach, and a detailed explanation is as follows:

1) The pre-trained RGB model is trained by hundreds of thousands of RGB face images, and the parameters of the model are acquired.
2) These parameters are set as the initial parameters of the NIR deep CNN.
3) Training of the NIR deep CNN model is conducted by finely updating the initial parameters of the model.

The above-mentioned fine-tuning approach alleviates the need for direct training. Direct training is a method where initial parameters of a deep CNN model are set randomly before training. The drawback of direct training is that recognition performance is degraded when using small NIR face databases. The proposed fine-tuning approach makes it possible to train the NIR deep CNN model effectively. This is due to the similarity between the parameters of the NIR deep CNN model and the pre-trained RGB model. As shown in Fig. 3, convolution filter outputs of the NIR deep CNN model and the pre-trained RGB model are highly similar. We used VGG-16 [1] as the deep CNN architecture of the NIR deep CNN model and the pre-trained RGB model. The two models were trained with direct training. In Fig. 3, the convolution layers' filter outputs for the pre-trained RGB model and the NIR deep CNN model were acquired from the same input NIR face image. The convolution filter outputs of the two models have the same characteristics, with high values in facial structures such as eyes, eyebrows, a nose, a mouth, and face contours. As a result, the convolution filters of the pre-trained RGB model and the NIR deep CNN model are similarly trained to detect the facial structures easily.

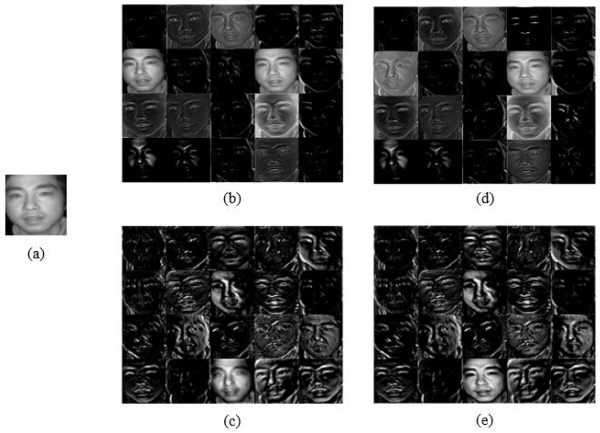

Fig. 3. Convolution filter outputs of the pre-trained RGB model and the NIR deep CNN model. (a) The input NIR face image. (b) and (c) The first and second convolution layers' filter outputs for the pre-trained RGB model, respectively. (e) and (f) The first and second convolution layers' filter outputs for the NIR deep CNN model, respectively.

2.2 Deep CNN Architectures for NIR FR

In this study, NIR FR was performed by off-the-shelf deep CNN architectures, and we confirmed the best one among the deep CNN architectures. The deep CNN architectures used for the performance evaluation are VGG-16 [1], Inception-Resnet-v1 [5], Inception-Resnet-v2 [5], DeepID2 [8], and ResNet-152 [4]. The size of the input NIR face images was 160x160, and the dimension of all extracted deep features from the input images was 128. Cross entropy loss was used to train all deep CNN architectures. In DeepID2, the deep CNN architecture was trained by cross-entropy loss and verification loss. Verification loss is similar to triplet loss, which was developed in FaceNet [10].

2.3 NIR Face Databases

In this study, we utilized several public NIR face databases to conduct the performance evaluation. The public databases are CASIA NIR [19], CASIA NIR-VIS 2.0 [20], PolyU-NIRFD [21], ND-NIVL [22] and INF databases. The INF database is constructed by integrating the CASIA NIR-VIS 2.0, PolyU-NIRFD, and ND-NIVL databases. To conduct performance evaluation for NIR FR in different environments from the public NIR face database, the VF_NIR, VF_RGB, VF_PLC_NIR, and VF_PLC_RGB databases were constructed using Intel® RealSense™ SR300. Fig. 4 shows sample images from the VF_NIR, VF_RGB, VF_PLC_NIR, and VF_PLC_RGB databases. VF_NIR and VF_RGB databases respectively contain the NIR and RGB face images, which were captured in an indoor fluorescent lighting environment. VF_PLC_NIR and VF_PLC_RGB databases were constructed in poor lighting conditions. The two databases include not only the face images in poor lighting conditions, but also those from the VF_NIR and VF_RGB databases. VF_NIR and VF_RGB database contain 500 NIR and RGB face images for 10 labels, respectively. In VF_PLC_NIR and VF_PLC_RGB databases, there are 550 NIR and RGB face images for 11 labels, respectively. All NIR face images were aligned based on MTCNN [17]. The effect of MTCNN is that eyes, noses, and mouths are aligned in similar positions on the images.

Table 1. The size of training and validation data in each public NIR face database

Database	# of training data	# of labels for training	# of validation data	# of labels for validation
CASIA NIR [19]	2,798	140	1,140	57
CASIA NIR-VIS 2.0 [20]	8,703	505	3,782	216
PolyU-NIRFD [21]	17,808	159	6,890	68
ND-NIVL [22]	18,314	403	3,667	165
INF	44,825	1,067	14,339	449

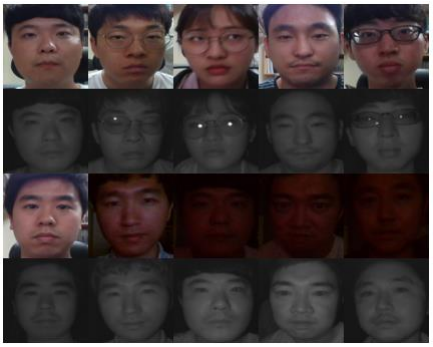

Fig. 4. RGB and NIR face images from Intel RealSense SR300 in various lighting conditions. VF_RGB, VF_NIR, VF_PLC_RGB and VF_PLC_NIR databases are in order from the first row.

3. EXPERIMENTAL RESULTS

In this section, we present the three experiments for showing the priority of the proposed fine-tuning approach. The three experiments are as follows.

3.1 Performances on Public NIR Face Databases

In this experiment, we showed the best NIR deep CNN architecture among off-the-shelf deep CNN architectures [1, 4, 5, 8] in the proposed fine-tuning approach. This experiment was conducted with the CASIA NIR [19], CASIA NIR-VIS 2.0 [20], PolyU-NIRFD [21], ND-NIVL [22], and INF databases. We split each public database into training and validation data. Table 1 shows the number of training and validation data for each public database along with the number of labels. To conduct a performance evaluation, 6000 validation pairs were extracted from validation data. Validation pairs contain 3000 positive and negative pairs, respectively.

Fig. 5 shows the validation rate of the NIR deep CNN architectures for the public NIR face databases [19-22]. The validation rate was calculated with a fixed FAR of 0.1%. VGG-16 [1], Inception-Resnet-v1 [5], and Inception-Resnet-v2 [5] achieved more than 99% performance on most public NIR face databases. VGG-16 showed the highest performances of 99.57%, 99.60%, and 99.57% on CASIA NIR-VIS 2.0 [20], ND-NIVL [22], and INF databases, respectively. Inception-Resnet-v1 achieved the highest performances of 99.27% and 99.70% on CASIA NIR [19] and PolyU-NIRFD [21] databases, respectively. DeepID2 [8] and ResNet-152 [4] showed lower NIR FR performances on all public NIR face databases than VGG-16, Inception-Resnet-v1, and Inception-Resnet-v2. In particular, the performances of these two NIR deep CNN architectures were lower than 90% on the CASIA NIR database. As a result, we showed that VGG-16 achieved the best performance in NIR FR based on the proposed fine-tuning approach, and Inception-Resnet-v1 and v2 showed similar performances to VGG-16.

3.2 Comparison with Existing Methods

In this experiment, we compared the performances of the proposed fine-tuning approach with those of the existing NIR FR methods [14, 15]. The performance evaluation was conducted for the identification scenario. VGG-16 [1], Inception-Resnet-v1 [5], and Inception-Resnet-v2 [5] were used as NIR deep CNN architectures in the proposed fine-tuning approach. Detailed information on databases is as follows:

Training images The training data of the INF database in Table 1 was used to train the NIR deep CNN model of the proposed fine-tuning approach and the existing methods [14, 15]. For the existing methods, the softmax classifier was trained. To construct training data for the classifier, we extracted 591 NIR face images from the CASIA NIR [19] database.

Gallery / Probe images In the fine-tuning approach, gallery and probe images are required to evaluate NIR FR performance. We used 591 NIR face images from the CASIA NIR [19] database as the gallery images. Then, all images from the CASIA NIR database other than the gallery images were used as the probe images. In the case of the existing methods [14, 15], the gallery images are not required for NIR FR due to the softmax classifier, and the probe images were used in the same manner as in the proposed fine-tuning approach.

The result of this experiment is shown in Table 2. The existing methods [14, 15] show identification rates of 90.89% and 88.65%, respectively. The existing methods cannot secure enough performance to be used as biometrics. In the proposed fine-tuning approach, we achieved identification rates of 98.15%, 97.22%, and 99.67% for VGG-16 [1], Inception-Resnet-v1 [5], and Inception-Resnet-v2 [5], respectively. This experiment was conducted in a poorly constrained environment. In other words, the training and probe images were acquired from different public databases. The proposed fine-tuning approach showed up to 11% higher performances than the existing methods. Considering the experimental results, the proposed fine-tuning approach secured better generalization ability in a poorly constrained environment than existing methods.

3.3 Comparison with RGB FR

To compare the generalization ability of the proposed approach with that of RGB FR, two experiments were conducted. NIR FR and RGB FR were trained based on the proposed fine-tuning appr-

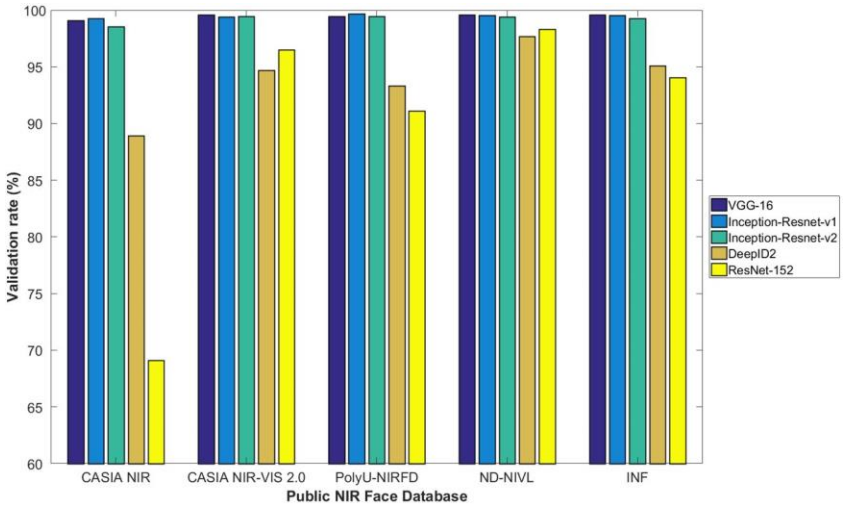

Fig. 5. The validation rate of the NIR deep CNN architectures according to the public NIR face databases.

Table 2. The identification rate of the proposed fine-tuning approach and existing methods.

NIR FR methods	Identification rate(%)
Zhang *et al*. [14]	90.89
Peng *et al*. [15]	88.65
Fine-tuning (VGG-16)	98.15
Fine-tuning (Inception-Resnet-v1)	97.22
Fine-tuning (Inception-Resnet-v2)	**99.67**

oach and RGB direct training, respectively. VGG-16 [1] was used as the deep CNN architecture for both NIR and RGB FR. A detailed explanation of these experiments is as follows:

Experiment 1 To compare the generalization ability of RGB and NIR FR in the poorly constrained environments, we assumed a real-world FR scenario. In this scenario, well constrained environments are insecure because many unexpected situations can occur during the FR process. In this experiment, to realize a real-world FR scenario, we constructed test databases in different environments from training and validation databases. The INF database was utilized for training and validation in NIR FR, and the VF_NIR database was used for testing. In the case of RGB FR, the CASIA WebFace [18], LFW [23], and VF_RGB databases were used for training, validation, and testing, respectively. The accuracy, validation rate, and FAR were used as the performance metrics.

Results of Experiment 1 As seen in Table 3, although RGB FR achieved a validation rate of 100.00%, a FAR is also high at 57.30%. For the FAR value, RGB FR cannot provide sufficient security to be used as a biometric in the poorly constrained environments. On the other hand, NIR FR based on the proposed approach achieved a FAR of 0.7%, and this value is considerably lower than for RGB FR. In addition, NIR FR achieved the high accuracy of 96.88% and high validation rate of 94.47%. As a result, NIR FR based on the proposed approach has better generalization ability than RGB FR for environmental variations.

Experiment 2 This experiment was conducted to compare the generalization ability of NIR and RGB FR in poor lighting conditions. The training methods and training databases of NIR and RGB FR were the same as in the first experiment. In this experiment, the VF_PLC_NIR database was used as the validation database for NIR FR. In the case of RGB FR, the VF_PLC_RGB

Table 3. The performances of the proposed fine-tuning approach and RGB FR in the real-world FR scenario.

Method	Accuracy(%)	Validation rate(%)	FAR(%)
NIR FR[a]	**96.88**	**94.47**	**0.7**
RGB FR	71.35	100.00	57.30

[a] NIR FR indicates the proposed fine-tuning approach.

Table 4. The performances of the proposed fine-tuning approach and RGB FR in the poor lighting conditions.

Method	Accuracy(%)	Validation rate(%)	FAR(%)
NIR FR[a]	**96.65**	**84.90**	**0.1%**
RGB FR	86.50	44.03	0.1%

database was utilized as the validation database. In this experiment, the validation database was used for performance evaluation. The accuracy and validation rate were used as performance metrics, and the FAR was fixed at 0.1%.

Results of Experiment 2 As shown in Table 4, RGB FR achieved an accuracy of 86.50% and validation rate of 44.03%. On the other hand, NIR FR achieved an accuracy of 96.65% and validation rate of 84.90%. The performances of NIR FR are significantly higher than those of RGB FR. Consequently, NIR FR based on the proposed fine-tuning approach has better generalization ability in the poor lighting conditions than RGB FR.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a fine-tuning approach to effectively train a deep CNN model for NIR FR by utilizing the parameters of a pre-trained RGB model. Also, we justified this fine-tuning approach by showing the similarity between the parameters of an NIR deep CNN model and a pre-trained RGB model. The proposed fine-tuning approach achieved high validation rates of more than 99% on the public NIR face databases. In addition, the proposed approach showed better generalization ability in various lighting conditions and environments than the existing NIR FR methods [14, 15] and RGB FR. However, we found that the FR performances of the proposed fine-tuning approach tend to be correlated with the type of NIR sensors. Therefore, in future work, we will focus on alleviating the sensor dependency of NIR FR.

5. REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," In *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1-14.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," In *Proc. CVPR*, Boston, MA, USA, 2015, pp. 1-9.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," In *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 2818-2826.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 770-778.

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," In *Proc. AAAI*, Phoenix, AZ, USA, 2016, pp. 4278-4284.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In *Proc. NIPS*, Stateline, NV, USA, 2012, pp. 1097-1105.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," In *Proc. BMVC*, Swansea, UK, 2015, pp. 1-12.

[8] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," In *Proc. NIPS*, Montreal, Canada, 2014, pp. 1988-1996.

[9] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," In *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 1701-1708.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," In *Proc. CVPR*, Boston, MA, USA, 2015, pp. 815-823.

[11] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," In *Proc. CVPR*, Boston, MA, USA, 2015, pp. 2892-2900.

[12] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015, February 3) DeepID3: Face recognition with very deep neural networks. Cornell University Library, NY. [Online]. Available: https://arxiv.org/pdf/1502.00873.pdf.

[13] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," In *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 6738-6746.

[14] X. Zhang, M. Peng, and T. Chen, "Face recognition from near-infrared images with convolutional neural network," In *Proc. WCSP*, Yangzhou, China, 2016, pp. 13-15.

[15] M. Peng, C. Wang, T. Chen, and G. Liu, "NIRFaceNet: A convolutional neural network for near-infrared face identification," *Information*, vol. 7, no. 4, pp.61-74, Oct. 2016.

[16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," In *Proc. NIPS*, Montreal, Canada, 2014, pp. 3320-3328.

[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.

[18] D. Yi, Z. Lei, S. Liao, and S. Z. Li. (2014, November 28) Learning face representation from scratch. Cornell University Library, NY. [Online]. Available: https://arxiv.org/pdf/1411.7923.pdf.

[19] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 627-639, Apr. 2007.

[20] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," In *Proc. CVPRW*, Portland, OR, USA, 2013, pp. 348-353.

[21] B. Zhang, L. Zhang, D. Zhang, and L. Shen, "Directional binary code with application to PolyU near-infrared face database," *Pattern Recognit. Lett.*, vol. 31, no. 14, pp. 2337-2344, Oct. 2010.

[22] J. Bernhard, J. Barr, K. W. Bowyer, and P. J. Flynn, "Near-IR to visible light face matching: Effectiveness of pre-processing options for commercial matchers," In *Proc. BTAS*, Arlington, VA, USA, 2015, pp. 1-8.

[23] 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," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning", *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[25] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting", In *Proc. NIPS*, Lake Tahoe, NV, USA, 2013, pp. 46-54.