Thermal Face Recognition based on Transformation by Residual U-Net and Pixel Shuffle Upsampling

Soumya Chatterjee and Wei-Ta Chu

[1] Indian Institute of Technology Bombay, India
soumyac1999@gmail.com
[2] National Cheng Kung University, Taiwan
wtchu@gs.ncku.edu.tw

Abstract. We present a thermal face recognition system that first transforms the given face in the thermal spectrum into the visible spectrum, and then recognizes the transformed face by matching it with the face gallery. To achieve high-fidelity transformation, the U-Net structure with a residual network backbone is developed for generating visible face images from thermal face images. Our work mainly improves upon previous works on the Nagoya University thermal face dataset. In the evaluation, we show that the rank-1 recognition accuracy can be improved by more than 10%. The improvement on visual quality of transformed faces is also measured in terms of PSNR (with 0.36 dB improvement) and SSIM (with 0.07 improvement).

Keywords: Thermal face recognition · thermal-to-visible transformation · thermal face verification.

1 Introduction

Thermal face recognition has been attracting more and more attention in the recent years due to its broad application in many domains like night-time surveillance and access control. Face recognition has been mainly focused on the visible spectrum but this depends on external conditions like illumination. Imaging in the visible spectrum involves measuring the light reflected by the face. Hence, changes in lighting conditions can cause significant changes in visual appearance and degrade the performance of such systems. Thermal infrared images are captured by passive infrared sensors which measure the radiations emitted by the facial tissues, and hence are independent of the external lighting.

Infrared images are categorized according to the wavelengths sensed, including near infrared 'NIR' ($0.74\mu - 1\mu$m), short-wave infrared 'SWIR' ($1\mu - 3\mu$m), mid-wave infrared 'MWIR' ($3\mu - 5\mu$m), and long-wave infrared 'LWIR' ($8\mu - 14\mu$m). NIR and SWIR imaging are reflection based and visual appearance of objects and are similar to visible images. Prior studies on NIR or SWIR images achieved promising recognition performance. On the contrary, MWIR and

LWIR images measure material emissivity and temperature. Skin tissue has high emissivity in both the MWIR and LWIR spectrums. Because of this natural difference between the reflective visible spectrum and sensed emissivity in the thermal spectrum, images taken in the two modalities are very different and have a large modality gap. This hinders reliable face matching across the visible spectrum and the MSIR/LWIR spectrums. Currently some studies have focused on MWIR and LWIR face images, but only limited performance have been achieved [14] [3] [5]. In the following, we would call the MWIR/LWIR spectrums as *thermal spectrum*, and call face images captured in the thermal spectrum as *thermal faces*.

(a) Thermal Image (b) Visible Image

Fig. 1: Examples from NU Dataset

The goal of thermal face recognition is to identify a person captured in the thermal spectrum by finding the most similar face images captured in visible spectrum. This task is thus a cross-modal matching problem, where we need a non-linear mapping from the thermal spectrum to the visible spectrum while preserving the identity information.

We present a deep convolutional neural network based on the U-Net [13] architecture for the thermal face recognition task. U-Nets have been widely used for various tasks including image segmentation [13], image feature extraction [2], etc. In our work, the U-Net is used to synthesize visible faces from given query thermal faces. The generated faces are used for matching against the gallery images. To improve upon visual quality of the generated visible face images, we propose a modified U-Net architecture using residual blocks instead of convolutional layers as the basic building components. It has been shown in [10] that the skip connections in residual networks [4] give rise to much smoother loss surfaces than similar networks without skip connections. Hence, they are easier to train and are able to find much better local optimums. In addition to this, we use pixel shuffle upsampling in the expansive part of our network in place of transposed convolutional layers. Pixel shuffle upsampling introduced in [15] achieves much better Peak Signal to Noise Ratio (PSNR) compared to other upsampling methods at a fraction of the computation cost.

In this paper, we evaluate the proposed networks on the thermal face dataset collected by Nagoya University [9] (which we will call the NU dataset). In the NU dataset, visible and thermal face pairs are available. In contrast to other thermal

face dataset, visible faces and thermal faces were captured simultaneously by two closely located cameras. Therefore, the visible face and the thermal face of the same individual are well aligned. Such alignment is important for us to clearly study the performance of our models.

The rest of this paper is organized as follows. Sec. 2 describes related works of thermal face recognition. Sec. 3 presents details of the proposed method. Evaluation results are shown in Sec. 4, followed by concluding remarks in Sec. 5.

2 Related Works

Existing thermal to visible face recognition works can be roughly grouped into two categories: (1) transforming faces in the thermal spectrum into the visible spectrum, and then conducting recognition; (2) projecting thermal faces and visible faces into the common feature space and then achieving recognition.

Kresnaraman et al. [9] transformed the given thermal face into a visible face by utilizing the relationship between images in the thermal and visible spectra obtained by canonical correlation analysis. Given a polarimetric thermal face that is composed of three channels, Riggan et al. [12] proposed to extract features and then estimate the corresponding visible face based on a regression model. Both feature extraction and regression are developed based on convolutional neural networks. The same research team later proposed to improve quality of face synthesis by jointly considering global (entire face) and local regions (eyes, nose, and mouth) [11]. With a similar idea, Chen and Ross [1] proposed to develop a semantic-guided generative adversarial network to transform thermal faces into visible faces. The semantic labels obtained by a face parsing network provide important clues to improve synthesis.

One of the pioneering works in the second category is the deep perceptual mapping (DPM) [14]. Sarfraz and Stiefelhagen first extracted dense SIFT features from patches of thermal faces, and then transformed these features by an auto-encoder. The objective of this auto-encoder is to map the given features into that similar to the features extracted from the corresponding visible face. Iranmanesh et al. [7] proposed a coupled deep neural network architecture to make full use of the polarimetric thermal information. Taking VGG-16 network as the basic building component, polarimetric thermal faces and corresponding visible faces are fed to VGG-16 like networks to extract and embed features.

Our work falls into the first category, and we mainly focus on basic thermal faces rather than polarimetric thermal faces. In the evaluation, we would compare several proposed variants with one from the first category, i.e., [9], and one from the second category, i.e., [14].

3 Methods

The thermal face recognition problem is formulated as follows. Assume that we have a set $\mathcal{G}$ of visible faces called the gallery set. Given a thermal face x, called the probe, the visible face in $\mathcal{G}$ which corresponds to the same person

as x needs to be found. For this, we design a mapping f from the domain of thermal face images $\mathcal{T}$ to visible face images $\mathcal{V}$. This mapping f is learnt from the training data using a deep convolutional neural network (CNN). Given a probe x, the visible face $\hat{y}$ is reconstructed using the learnt mapping function f, i.e., $\hat{y} = f(x)$. The Euclidean distance between $\hat{y}$ and each of the images in $\mathcal{G}$ is calculated, and the one with minimum distance to $\hat{y}$ is returned as the match y^*. That is,

$$y^* = \underset{t \in \mathcal{G}}{\operatorname{argmin}} \left\| t - f(x) \right\|_2 . \tag{1}$$

Alternatively, a pretrained face recognition network can be used to find the best match. In this case, the pretrained network can be thought of a mapping g from the domain of visible face images V to $\mathbb{R}^N$, i.e., given a visible face images, the network gives an encoding of the input as a vector of size N. The encodings are such that Euclidean distance between encodings of images of the same person is small while it is large for different people. For this,

$$y^* = \underset{t \in \mathcal{G}}{\operatorname{argmin}} \left\| g(t) - g\left(f(x)\right) \right\|_2 . \tag{2}$$

3.1 U-Net Model for Face Recognition

The transformation function f that transforms a thermal face into a visible face is the most critical component in this work. We use a U-Net to develop the function f. Figure 2 shows the network structure. A U-Net has a contracting path and an expansive path. The contracting path is a conventional convolutional neural network, where the convolutional kernel is 3×3 with stride 2 and same padding; the activation function is ReLU, followed by a 2×2 max pooling for downsampling.

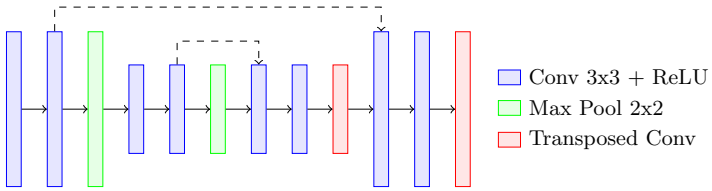


Fig. 2: The baseline network architecture. The dashed lines represent the skip connections in U-Net. Notice that the baseline encoder-decoder model does not have these connections.

Following [13], the number of feature channels are doubled at each downsampling step. In the expansive path we halve the number of feature channels at each upsampling step, by 2×2 transposed convolution. Every step in the expansive path contains upsampling, a concatenation with the corresponding feature map from the contracting path, and two convolutional layers similar to those in the

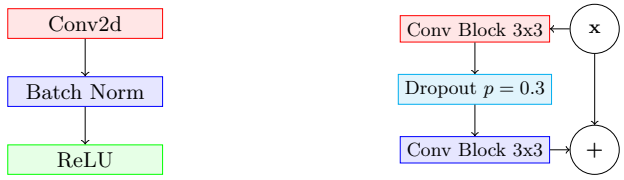
contracting path. The last 1×1 convolution used in [13] has been omitted in our work, since we found that it gives better results. To demonstrate effectiveness of the skip connection in U-Net architecture, we will compare our architecture with a baseline model without these skip connections. Both the networks are trained using the same protocol based on a weighted combination of mean square error (MSE) loss and perceptual loss, which will be described later.

3.2 ResNet U-Net with Pixel Shuffle

The expansive part of the U-Net is implemented by deconvolution and conducts a sequence of upsampling. Given the feature maps generated by the contracting part, we can view the expansive part as a sequence of super-resolution processes, which upsample a low-resolution image into a high-resolution one. Shi et al. [15] proposed an efficient sub-pixel convolutional neural network to enable real-time super-resolution. Not only computational efficiency, the proposed sub-pixel CNN also yields better visual quality of high-resolution images.

In our work, we propose to view the expansive part of a U-Net as the process of super-resolution, and attempt to integrate sub-pixel CNNs to get performance improvement. This method is usually also called *pixel shuffle*, and we will take this term in the following. In addition, we further try to consider residual blocks [4] to improve visual quality of transformed faces.

Pixel Shuffle Upsampling: Upsampling by a factor r can be achieved by transposed convolution with a stride r or a fractionally strided convolution with a stride $\frac{1}{r}$ [15]. Pixel shuffle upsampling is an efficient implementation of the fractionally strided convolution. In this a $H \times W \times C \cdot r^2$ tensor is rearranged into a $rH \times rW \times C$ tensor thus achieving an upsampling by a factor r. Our network uses $r = 2$ to upsample tensors by a factor of 2 in the expansive part.



(a) Convolutional block (b) Residual block

Fig. 3: Illustration of convolutional blocks and residual blocks.

Residual Blocks: To further improve upon visual quality of the generated visible face images, we make the following modifications to our above network. The convolutional layers used are replaced by residual blocks each consisting of two convolutional blocks with a dropout [17] layer in between, as shown in Figure 3b. Each convolutional block (Figure 3a) is made up of a 3×3 convolutional layer, and a batch normalization layer [6] followed by ReLU activation. In the residual

blocks, using ideas from [19], the number of input channels is first doubled and then halved such that the resultant tensor has the same shape as the input. We use 2×2 max pooling for downsampling but in the upsampling path, pixel shuffle upsampling [15] is used instead of transposed convolution. The number of channels are doubled and halved in the downsampling and upsampling, respectively, using 1×1 convolutional layers following each residual block. The final network architecture is shown in Figure 4.

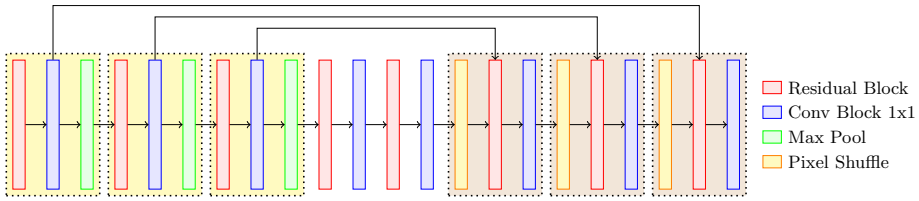


Fig. 4: The architecture of the proposed U-Net with residual blocks and pixel shuffle upsampling.

3.3 Losses

We use a weighted combination of two loss functions, namely mean squared error loss and perceptual loss to train our networks. These loss functions are described below.

Perceptual Loss proposed in [8] is used to measure the high-level semantic differences between transformed images and target images. It ensures that the transformed image is perceptually similar to the target. Given a transformed face $\hat{y}$ and a visible face y, a VGG-19 network [16] pretrained on the ImageNet dataset is used to extract features from them. The feature maps output by the last layer of each convolutional block are taken as the features. Let $\phi_j(\hat{y})$ and $\phi_j(y)$ denote the features extracted by the j considered convolutional layer, from the transformed face $\hat{y}$ and the visible face y, respectively. Perceptual loss between $\phi_j(\hat{y})$ and $\phi_j(y)$, both of shape $H_j \times W_j \times C_j$, is defined as

$$\mathcal{L}_j^{feat}(\hat{y}, y) = \frac{1}{H_j W_j C_j} \left\| \phi_j(\hat{y}) - \phi_j(y) \right\|_2^2, \tag{3}$$

where H_j and W_j denote height and width of a feature map $\phi_j(\hat{y})$ (or $\phi_j(y)$), and C_j denotes the number of channels (feature maps). Overall, the perceptual loss between an image and its reconstruction is the mean of $\mathcal{L}_j^{feat}$ over all j.

Mean Squared Error Loss ensures that the identity of the thermal face is preserved in the reconstructed visible face, i.e., ensuring fidelity between the thermal images and its transformation.

$$\mathcal{L}^{MSE}(\hat{y}, y) = \frac{1}{HW} \left\| \hat{y} - y \right\|_2^2. \tag{4}$$

For training, we linearly combine two losses:

$$\mathcal{L}(\hat{y}, y) = \mathcal{L}^{MSE}(\hat{y}, y) + \lambda \cdot \mathcal{L}^{feat}(\hat{y}, y), \tag{5}$$

where $\lambda = 0.01$.

4 Results

4.1 Evaluation Dataset

We evaluate the proposed networks on the thermal face dataset from Nagoya University, which consists of 180 Japanese subjects (169 males and 11 females). Five pairs of thermal faces and visible faces were captured for each individual. The ordinary camera capturing the visible spectrum and the thermal camera capturing LWIR images were mounted closely, and the same pair of thermal and visible faces were captured simultaneously, making the image pairs very well aligned. This characteristic is distinct to other thermal face datasets, and provides us a good foundation for the proposed study. There are thus 900 thermal images and 900 visible images in total. All of them are frontal faces with neutral expression. The thermal images were captured by the Advanced Thermo TVS-500EX camera, which senses wavelength ranging from 8μm to 14μm. The corresponding thermal and visible images were captured at the same time, and underwent the same preprocessing. After cropping, calibration, and resizing, resolution of both types of images is 56×64 pixels. Fig. 1 shows a sample image pair from the NU database.

In the evaluation protocol of [9] and [3], 180 individuals are separated into two parts, i.e., 160 people and 20 people. The 160 people in the first part are equally divided into 16 groups, i.e., each group consists of 10 people. Among the 16 groups, 15 groups are selected as the training set. Thermal faces of the remaining group, consisting of 10 people, are taken as the probe image set (test data). In the gallery set, in addition to visible images corresponding to these 10 people, the 20 people in the second part separated at the beginning are also included in the gallery set to increase the number of candidate identities, i.e., increasing noise. These 20 people are also used as the validation set during training.

The models are trained with the Adam optimizer with learning rate of 0.01 for 150 epochs. A learning rate decay of 0.6 every 25 epochs is also used.

4.2 Face Recognition

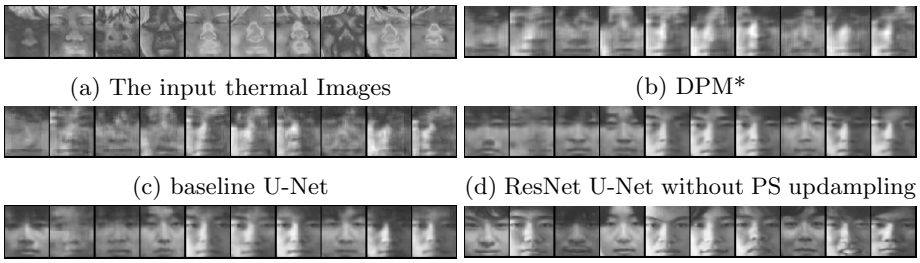
First, we transform the thermal face images from the probe set to visible spectrum using the model illustrated in Fig. 4. The transformed images are then matched against the visible face images in the gallery set, and the one with the minimum Euclidean distance from the transformed probe is selected. The match is considered to be correct if the selected visible face and the query thermal face are of the same individual. We measure the rank-1 recognition

accuracy which is averaged over 5 different splits of the dataset. The results of canonical correlation analysis (CCA) [9], deep perceptual model (DPM) [3] and the baseline encoder-decoder model (illustrated in Fig. 2) are compared with the proposed model (ResNet U-Net without/with Pixel Shuffle, illustrated in Fig. 4), as shown in Table 1. Please notice that all the methods but DPM in this table are developed for transforming thermal faces into visible faces. DPM uses a basic auto-encoder to transform features extracted from thermal faces into that similar to the corresponding visible faces. We take the auto-encoder part as the transformation, and do the same thing as other methods. Because this is not the original DPM, we denote this approach DPM* in Table 1, Table 2, and Figure 5.

Table 1: Rank-1 thermal face recognition accuracy.

Methods	Average Accuracy (%)
CCA [9]	14.00
DPM* [3]	59.50
Baseline U-Net	67.60
ResNet U-Net (w/o PS)	68.80
ResNet U-Net (w. PS)	**69.60**

As can be seen, the baseline U-Net improves recognition accuracy over the basic auto-encoder from 59.50% to 67.60%. The proposed ResNet U-Net with pixel shuffle further provides performance over the baseline U-Net by 2%. This result shows effectiveness of taking residual blocks as the components in the U-Net over the baseline U-Net. The skip connections allow the model to infer finer details which may be lost in the downsampling part of the network.



(a) The input thermal Images (b) DPM*

(c) baseline U-Net (d) ResNet U-Net without PS updampling

(e) ResNet U-Net with PS upsampling (f) Ground Truths

Fig. 5: Comparison of reconstructed visible images for all our models

4.3 Face Transformation

The main idea of this work is to transform thermal faces into visible faces, and then conduct face recognition in the visible domain. Therefore, we would like to evaluate the quality of transformed faces in the following.

We compare transformed faces by DPM [3], the baseline U-Net model, the ResNet U-Net model without pixel shuffle (PS) upsampling, and the ResNet U-Net model with PS upsampling. Fig. 5 shows sample transformation results of different models. From these samples, we clearly can see that DPM and the baseline U-Net yield relatively blurry visible faces. On the other hand, comparing Fig. 5c with Fig. 5d, the ResNet U-Net model without PS upsampling yields much clearer faces, which shows the effectiveness of residual blocks. The ResNet U-Net model with PS upsampling further slightly improves visual quality.

Table 2: Visual quality comparison of different transformation methods.

Method	PSNR (dB)	SSIM
CCA [9]	20.260	0.730
DPM* [3]	19.709	0.705
baseline U-Net	19.499	0.672
ResNet U-Net (w/o PS)	19.803	0.781
ResNet U-Net (w. PS)	**20.627**	**0.803**

To quantitatively evaluate different models, we calculate the average Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) [18] values of transformed faces. Table 2 lists the values. Two observations can be made from this table. First, values of PSNR and SSIM may not always be consistent with human perception. For example, samples in Fig. 5d look much better than Fig. 5b, but in terms of PSNR, the ResNet U-Net model without PS upsampling (PSNR=19.803) is just slightly better than the DPM (PSNR=19.709). Second, our ResNet U-Net model with PS upsampling consistently performs better than all other methods on both the metrics. Also, the results with PS upsampling are better than those without it, which justifies the effectiveness of viewing a part of transformation as a super-resolution process.

5 Conclusion

We have presented a framework to reconstruct visible faces from thermal faces preserving identity of the subject. In the proposed network, a U-Net model with residual blocks as the building components are used. Using sub-pixel convolution and pixel shuffle upsampling, we are able to transform thermal faces into realistic visible faces. Based on the transformed faces, recognition performance better than the canonical deep perceptual model and other variants can be obtained. In the future, more extensive experiments can be conducted. We would build

a robust thermal face detection system so that these combined can be used for thermal face recognition in real-world situation.

6 Acknowledgement

This work was partially supported by the Ministry of Science and Technology under the grant 108-2221-E-006-227-MY3, 107-2221-E-006-239-MY2, 107-2923-E-194-003-MY3, 107-2627-H-155-001, and 107-2218-E-002-055.

References

1. Chen, C., Ross, A.: Matching thermal to visible face images using a semantic-guided generative adversarial network. In: Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition (2019)
2. Chu, W.T., Liu, Y.H.: Thermal facial landmark detection by deep multi-task learning. In: Proceedings of IEEE International Workshop on Multimedia Signal Processing. IEEE (2019)
3. Chu, W.T., Wu, J.N.: A parametric study of deep perceptual model on visible to thermal face recognition. In: Proceedings of IEEE Visual Communications and Image Processing (2018)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (2016)
5. Hu, S., Choi, J., Chan, A.L., Schwartz, W.R.: Thermal-to-visible face recognition using partial least squares. Journal of the Optical Society of America A **32**(3), 431–442 (2015)
6. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning. pp. 448–456 (2015)
7. Iranmanesh, S.M., Dabouei, A., Kazemi, H., Nasrabadi, N.M.: Deep cross polarimetric thermal-to-visible face recognition. In: Proceedings of International Conference on Biometrics (2018)
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of European Conference on Computer Vision. Springer (2016)
9. Kresnaraman, B., Deguchi, D., Takahashi, T., Mekada, Y., Ide, I., Murase, H.: Reconstructing face image from the thermal infrared spectrum to the visible spectrum. Sensors **16**(4) (2016)
10. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. In: Proceedings of Advances in Neural Information Processing Systems (2018)
11. Riggan, B.S., Short, N.J., Hu, S.: Thermal to visible synthesis of face images using multiple regions. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (2018)
12. Riggan, B.S., Short, N.J., Hu, S., Kwon, H.: Estimation of visible spectrum faces from polarimetric thermal faces. In: Proceedings of IEEE International Conference on Biometrics Theory, Applications and Systems (2016)

13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
14. Sarfraz, M.S., Stiefelhagen, R.: Deep perceptual mapping for thermal to visible face recognition. International Journal of Computer Vision **122**(3), 426–438 (2017)
15. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (2016)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations (2015)
17. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**, 1929–1958 (2014)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
19. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)