Thermal Face Recognition and Thermal to Visible Face Reconstruction using U-Nets

Soumya Chatterjee
IIT Bombay

Wei-Ta Chu
National Chung Cheng University

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).

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\(\mu\)m – 1\(\mu\)m), short-wave infrared ‘SWIR’ (1\(\mu\)m – 3\(\mu\)m), mid-wave infrared ‘MWIR’ (3\(\mu\)m – 5\(\mu\)m), and long-wave infrared ‘LWIR’ (8\(\mu\)m – 14\(\mu\)m). NIR and SWIR imaging are reflection based and their visual appearance 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 spectrum. 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 two domains. Currently some studies have focused on MWIR and LWIR face images but only limited performance have been achieved [1] [2] [3].

The goal of thermal face recognition is to identify a person captured in infrared 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 infrared spectrum to visible spectrum while preserving the identity information.

We present a Deep CNN model based on the U-Net [4] architecture for the thermal face recognition task. U-Nets have been widely used for various tasks including image segmentation [4], generators in GANs, etc. The U-Net is used to synthesize visible face images from the corresponding thermal images. The generated faces are used for matching against the gallery images. To improve upon the 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 [5] that the skip connections in residual networks [6] 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 \cite{7} 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 our networks on the thermal face dataset collected by Nagoya University \cite{8} (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.

2 Methods

First, we formulate the tasks. In thermal face recognition, 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 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 of \( \hat{y} \) is calculated with all the images in \( \mathcal{G} \) and the one with minimum distance is returned as the match \( y \), i.e.

\[
y = \arg\min_{t \in \mathcal{G}} \| t - f(x) \|_2
\]

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 \( \mathcal{V} \) 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,

\[
y = \arg\min_{t \in \mathcal{G}} \| g(t) - g(f(x)) \|_2
\]

In face verification, we are given two images \( x_1 \) and \( x_2 \) and we have to decide whether they belong to the same person. For this we can use a pretrained network similar to face recognition except that here the only condition is that the encodings of the same person have to similar. We say that \( x_1 \) and \( x_2 \) are the same person if \( \| g(t) - g(f(x)) \|_2 < \epsilon \) for some suitable \( \epsilon \).

2.1 U-Net Model for Face Recognition

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 \( 3 \times 3 \) with stride 2 and same padding; the activation function is ReLU, followed by a \( 2 \times 2 \) max pooling for downsampling.

Following \cite{4}, 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 \( 2 \times 2 \) 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 \( 1 \times 1 \) convolution used in \cite{4} has been omitted since we found that it gives better results. To demonstrate the effectiveness of the skip connection in U-Net architecture, we compare our architecture with a baseline model without these skip connections. Both the networks are trained using the same protocol using a weighted combination of MSE loss and Perceptual loss.

2.2 ResNet U-Net with Pixel Shuffle

In this subsection, we first describe Pixel Shuffle upsampling followed by description of our model.

**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 \( \frac{r}{2} \). Pixel shuffle upsampling is an efficient implementation of the fractionally strided convolution. In this a \( H \times W \times C \) \( \times r^2 \) tensor is rearranged into a \( rH \times rW \times C \) tensor thus achieving an upsampling by a factor \( r \) (Figure 3).
Our network uses $r = 2$ to upsample tensors by a factor of 2 in the expansive part.

![Building Blocks](image)

Figure 4: Building Blocks

To further improve upon the 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 layer in between as shown in Figure 4b. Each convolutional block (Figure 4a) is made up of a $3 \times 3$ convolutional layer, a batch normalization layer followed ReLU activation. In the residual blocks, using ideas from [10], the number of inputs channels is first doubled and then halved such that the resultant tensor has the same shape as the input. We use $2 \times 2$ max pooling for downsampling but in the upsampling path, Pixel Shuffle upsampling [7] is used instead of transposed convolution. The number of channels are doubled and halved in the downsampling and upsampling respectively using $1 \times 1$ convolutional layers following each residual block. The final network architecture is shown in Figure 4.

### 2.3 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 by [11] is used to measure the high-level semantic differences between the synthesized and target images. It ensures that the synthesized image is perceptually similar to the target. We use features $\phi_j$ extracted from different layers of a VGG-19 network pretrained on ImageNet dataset for calculating the it. Perceptual loss (feature reconstruction loss) between a feature map $y$ of shape $H_j \times W_j \times C_j$ extracted from layer $j$ and its reconstruction $\hat{y}$ is defined as

$$
\mathcal{L}_{j}^{feat}(\hat{y}, y) = \frac{1}{H_jW_jC_j} \| \phi_j(\hat{y}) - \phi_j(y) \|^2_2
$$

and 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., ensures fidelity between the thermal images and its reconstruction.

$$
\mathcal{L}^{MSE}(\hat{y}, y) = \frac{1}{HWC} \| \hat{y} - y \|^2_2
$$

For training, we use

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

where $\lambda = 0.01$.

### 3 Results

#### 3.1 Evaluation Dataset

We evaluate our networks on the thermal face dataset from Nagoya University consisting of 180 Japanese subjects ($169$ males and $11$ females). Five pairs of thermal images and visible image were captured for each individual. The ordinary camera capturing visible image and the infra-red camera capturing LWIR images were mounted closely, and the same pair of thermal and visible images 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 ranged from (8 mm – 14 mm). 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 x 64 pixels. Fig 1 shows a sample image pair from the NU database.

We attempt to verify effectiveness of our frameworks on the NU database. In the evaluation protocol of [8] and [2], 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 for face recognition are trained using Adam optimizer [12] with learning rate of 0.001 for 100 epochs. The models for thermal to visible face reconstruction are trained with Adam optimizer with learning rate of 0.01 for 150 epochs. A learning rate decay of 0.6 every 25 epochs is also used.
3.2 Face Recognition

First, we transform the thermal face images from the probe set to visible spectrum using our model. 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 they 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 CCA [8], DPM [2] and our baseline encoder-decoder model is compared with our U-Net based model in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg. Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>14.00</td>
</tr>
<tr>
<td>DPM</td>
<td>59.50</td>
</tr>
<tr>
<td>Ours Enc-Dec</td>
<td>67.60</td>
</tr>
<tr>
<td>Ours U-Net</td>
<td>69.60</td>
</tr>
</tbody>
</table>

Table 1: Rank-1 Recognition Accuracy

Our model preforms better than the previous works on the NU dataset improving upon the accuracy by more that 10%. The results also show the effectiveness of the skip connections in the U-Net over the simple encoder-decoder model. The skip connections allow the model to infer finer details which may be lost in the downsampling part of the network hence improving accuracy by 2%.

3.3 Reconstruction

3.3.1 Evaluation Metrics

Peak Signal to Noise Ratio (PSNR) It is used to measure the quality of reconstruction of images. Higher PSNR indicates better quality of reconstructed images. PSNR between an image $I$ of size $m \times n$ with maximum possible pixel intensity $MAX_I(= 1)$ and its noisy reconstruction $K$ is defined as

$$PSNR = 10 \cdot \log_{10} \left( \frac{MAX_I^2}{MSE} \right) = -10 \cdot \log_{10}(MSE)$$

where

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |I(i, j) - K(i, j)|^2$$

Structural Similarity (SSIM) This metric introduced in [13] compares the luminance, contrast and structure of images. SSIM between an image $x$ with mean $\mu_x$ and variance $\sigma_x^2$ and another image $y$ with mean $\mu_y$ and variance $\sigma_y^2$ is defined as below. $\sigma_{xy}$ is the covariance of $x$ and $y$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)}$$

c_1, c_2 are constants to stabilize division

3.3.2 Evaluation Results

We compare the PSNR and SSIM values for CCA [2], our encoder-decoder model, our U-Net model, our ResNet U-Net model with Nearest Neighbour (NN) and Pixel Shuffle (PS) upsampling.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR (dB)</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>20.260</td>
<td>0.730</td>
</tr>
<tr>
<td>Encoder Decoder</td>
<td>19.709</td>
<td>0.705</td>
</tr>
<tr>
<td>U-Net</td>
<td>19.499</td>
<td>0.672</td>
</tr>
<tr>
<td>ResNet U-Net (NN)</td>
<td>19.803</td>
<td>0.781</td>
</tr>
<tr>
<td>ResNet U-Net (PS)</td>
<td>20.627</td>
<td><strong>0.803</strong></td>
</tr>
</tbody>
</table>

Table 2: Visual Appearance Comparison

It can be seen from Table 2 that our ResNet U-Net model with Sub-Pixel convolution and Pixel Shuffle upsampling performs better than all other methods on both the metrics. Also, the results with Pixel Shuffle upsampling are better than those with Nearest Neighbour upsampling justifying its effectiveness.

4 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 models with residual blocks as the building components are used. Using Sub-Pixel convolution and Pixel Shuffle upsampling, we are able to reconstruct realistic visible images. We also use another U-Net model for thermal face recognition using the reconstructed images. In the future, we want to build a robust thermal face detection so that these combined can be used for thermal face recognition in real-world situation.

Acknowledgements

This research was supported by grants from Advanced Institute of Manufacturing with High-tech Innovations, National Chung Cheng University and the Ministry of Education, Taiwan, Republic of China.

References