Deep Perceptual Mapping for Thermal to Visible Face Recognition

M. Saquib Sarfraz
https://cvhci.anthropomatik.kit.edu/~ssarfraz
Rainer Stiefelhagen
http://cvhci.anthropomatik.kit.edu

Institute of Anthropomatics & Robotics
Karlsruhe institute of Technology
Karlsruhe, Germany.

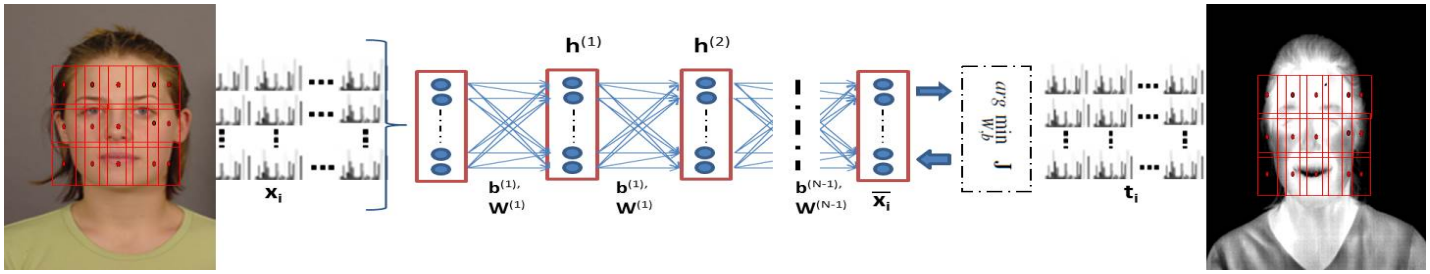

Figure 1: Deep Perceptual Mapping (DPM): densely computed features from the visible domain are mapped through the learned DPM network to the corresponding thermal domain.

Cross modal face matching between the thermal and visible spectrum is a much desired capability for night-time surveillance and security applications. Due to a very large modality gap, thermal-to-visible face recognition is one of the most challenging face matching problem. In this paper, we present an approach to bridge this modality gap by a significant margin. Our approach captures the highly non-linear relationship between the two modalities by using a deep neural network. Our model attempts to learn a non-linear mapping from visible to thermal spectrum while preserving the identity information. We show substantive performance improvement on a difficult thermal-visible face dataset (UND-X1). The presented approach improves the state-of-the-art by more than 10% in terms of Rank-1 identification and bridge the drop in performance due to the modality gap by more than 40%.

The goal of training the deep network is to learn the projections that can be used to bring the two modalities together. Typically, this would mean regressing the representation from one modality towards the other.

We construct a deep network comprising $N+1$ layers with $m^{(k)}$ units in the k-th layer, where $k = 1, 2, \cdots, N$. For an input of $x \in \mathbb{R}^d$, each layer will output a non-linear projection by using the learned projection matrix $\mathbf{W}$ and the non-linear activation function $g(\cdot)$. The output of the k-th hidden layer is $h^{(k)} = g(\mathbf{W}^{(k)} h^{(k-1)} + \mathbf{b}^{(k)})$, where $\mathbf{W}^{(k)} \in \mathbb{R}^{m^{(k)} \times m(k-1)}$ is the projection matrix to be learned in that layer, $\mathbf{b}^{(k)} \in \mathbb{R}^{m^{(k)}}$ is a bias vector and $g : \mathbb{R}^{m^{(k)}} \mapsto \mathbb{R}^{m^{(k)}}$ is the non-linear activation function. Similarly, the output of the most top level hidden layer can be computed as:

$$\mathbf{H}(x) = h^{(N)} = g(\mathbf{W}^{(N)} h^{(N-1)} + \mathbf{b}^{(N)}) \quad (1)$$

where the mapping $\mathbf{H} : \mathbb{R}^d \mapsto \mathbb{R}^{m^{(N)}}$ is a parametric non-linear perceptual mapping function learned by the parameters $\mathbf{W}$ and $\mathbf{b}$ over all the network layers. To determine the parameters $\mathbf{W}$ and $\mathbf{b}$ for such a mapping, our objective function must seek to minimize the perceptual difference between the visible and thermal training examples in the least mean square sense. We, therefore, formulate the DPM learning as the following optimization problem.

$$arg \min_{W,b} \quad \mathbf{J} = \frac{1}{M} \sum_{i=1}^{M} (\bar{x}_i - t_i)^2 + \frac{\lambda}{N} \sum_{k=1}^{N} (\|\mathbf{W}^{(k)}\|_F^2 + \|\mathbf{b}^{(k)}\|_2^2) \quad (2)$$

The first term in the objective function corresponds to the simple squared loss between the network output $\bar{x}$ given the visible domain input and the corresponding training example t from the thermal domain. The second term in the objective is the regularization term with λ as the regularization parameter. $\|\mathbf{W}\|_F$ is the Frobenius norm of the projection matrix $\mathbf{W}$. Given a training set $\mathbf{X} = \{x_1, x_2, \cdots, x_M\}$ and $\mathbf{T} = \{t_1, t_2, \cdots, t_M\}$ from visible and thermal domains respectively, the objective of training is to minimize the function in equation 2 with respect to the parameters $\mathbf{W}$ and $\mathbf{b}$.

The network is trained on densely computed feature representations (SIFT vectors) from overlapping small regions in the images. This proves effective, as the model is able to capture the differing local region's perceptual differences well. The training set comprises of these vectors coming from the corresponding patches from the images of the same identity. Using the corresponding images of the same identity ensures that the model will learn the only present differences due to the modality. Figure 1 capsulizes this process.

After obtaining the mapping from visible to thermal domain, we can now pose the matching problem as that of comparing the thermal images with that of mapped visible data. The presented set-up is ideal for the surveillance scenario as the gallery images can be processed and stored offline while at test time no transformation and overhead is necessary. We use a simple matrix vector multiplication to compute the similarity enabling us to match the probes in real-time.

We report evaluations using typical identification and verification settings. As baseline we use the same concatenated SIFT features but without the DPM mapping. This enables us to directly compare and see the effectiveness of the proposed model. Our results show (see paper) that we improve the state-of-the-art best published results by more than 10% in all cases.

Effect of Modality gap: Performance with 1 Gallery image/subject			
Thermal-Thermal	Thermal-Visible	Thermal-Visible (via DPM)	Modality-gap bridged
89.47	30.36	55.36	$\sim 42\%$

Table 1: Performance drop due to Modality gap: Rank-1 identification using 1 image/subject as gallery in Thermal-Thermal and Thermal-Visible matching using baseline features.

Effect of modality gap: We also present the experiment to measure the effect of modality gap. Keeping everything fixed *i.e.* using the same baseline features and settings, we compute the rank-1 identification score within the same modality. As shown in Table 1, we obtain 89.7% rank-1 score in the Thermal-Thermal identification scenario and 30.3% in the corresponding Thermal-visible scenario (using the same baseline features). This amounts to the performance drop, purely due to modality change, of about 59%. This reflects the challenging nature of the problem and the existing research gap to tackle this. As shown, with DPM on the same features, the performance is improved by 25%. This amounts to bridging the existing modality gap of 59% by more than 40%.

Computational Time: Training the DPM on 12 cores 3.2-GHz CPU takes between $1-1.5$ hours on MATLAB. Preprocessing, features extraction and mapping using DPM only takes $45ms$ for one image. This is even less in the testing case since no mapping is required for thermal images. It is, therefore, very fast and capable of running in real-time at ~ 28 fps.

Conclusively, the presented DPM approach is very effective, easy to train, real-time capable and provides a practical solution for a large surveillance and military application industry.