CrossMark

Deep Perceptual Mapping for Cross-Modal Face Recognition

M. Saquib Sarfraz[1] · **Rainer Stiefelhagen**[1]

Abstract 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 the visible to the thermal spectrum while preserving the identity information. We show substantive performance improvement on three difficult thermal–visible face datasets. The presented approach improves the state-of-the-art by more than 10 % on the UND-X1 dataset and by more than 15–30 % on the NVESD dataset in terms of Rank-1 identification. Our method bridges the drop in performance due to the modality gap by more than 40 %.

1 Introduction

Face recognition, mainly, has been focused in the visible spectrum. This pertains to a large number of applications from biometrics, access control systems, social media tagging to person retrieval in multimedia. Among the main

✉ M. Saquib Sarfraz
   saquib.sarfraz@kit.edu

[1] Institute of Anthropomatics & Robotics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

challenges in visible face recognition, the different lighting/illumination condition has proven to be one of the big factors for appearance change and performance degradation. Many prior studies such as Li et al. (2007), Socolinsky and Selinger (2002), Nicolo and Schmid (2012), and Klare and Jain (2013) have stated better face recognition performance in the infra-red spectrum because it is invariant to ambient lighting. Relatively recently, few efforts have been devoted in the cross-modal face recognition scenarios, where the objective is to identify a person captured in infra-red spectrum based on its stored high resolution visible face image. The motivation for this lies in the night-time or low light surveillance tasks where the image is captured discretely or covertly through active or passive infra-red sensors. In fact there does not exist a reliable solution for matching thermal faces, acquired covertly in such conditions, against the stored visible database such as police mugshots. This poses a significant research gap for such applications in aiding the law enforcement agencies. In the infra-red spectrum, thermal signatures emitted by skin tissues can be acquired through passive thermal sensors without using any active light source. This makes it an ideal candidate for covert night-time surveillance tasks.

As opposed to the visible spectrum wavelength (0.35–0.74 µm), the infra-red spectrum lies in four main ranges. Near infra-red 'NIR' (0.74–1 µm), short-wave infra-red 'SWIR' (1–3 µm), mid-wave infra-red 'MWIR' (3–5 µm) and long-wave infra-red 'LWIR' (8–14 µm). Since the NIR and SWIR bands are reflection dominant, they are more close to the visible spectrum. The MWIR and LWIR are the thermal spectrum and are emission dominant i.e. dependent on 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
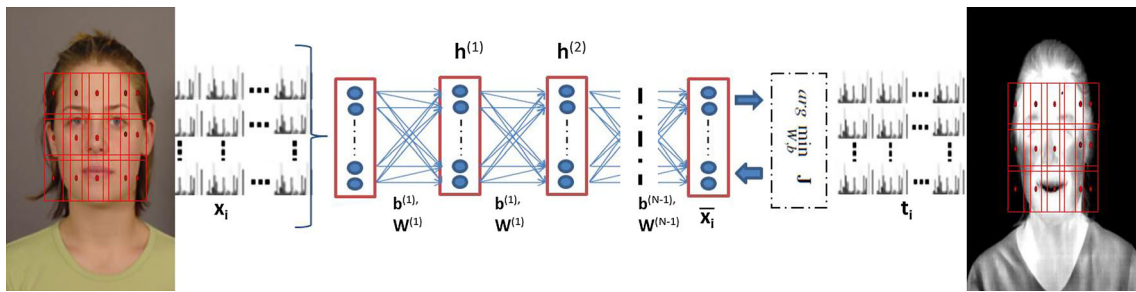
Fig. 1 Deep perceptual mapping (DPM): densely computed features from the visible domain are mapped through the learned DPM network to the corresponding thermal domain

taken in the two modalities are very different and have a large modality gap. This hinders reliable face matching across the two domains. It is, perhaps, for this reason that most of the earlier studies, aiming at cross-modal face recognition, rely only on visible-to-NIR face matching. While achieving very good results, NIR imaging use an active light source that makes it redundant for covert night-time surveillance. More recently, some attempts have been made in thermal-to-visible face recognition, indicating a significant performance gap due to the very challenging nature of the problem and the large modality gap.

In this paper, we seek to bridge this gap by trying to directly model the highly non-linear mapping between the two modalities. Our contribution is a useful model, based on a feed-forward deep neural network, and its effective design steps in order to map the perceptual differences between the two modalities while preserving the identity information. We show that this mapping can be learned from relatively little training data and that it works quite well in practice. Our model tries to learn a non-linear regression function between the visible and thermal data, where this data is comprised of densely pooled feature vectors from the images in the corresponding domain. Figure 1 summarises our approach. The learned projection matrices capture the non-linear relationship well and are able to bring the two closer to each other. Another contribution is to further the best published state-of-the-art performance on very challenging datasets i.e., the University of Notre Dame's 'UND-X1' LWIR-visible dataset by more than 10 % and on Night Vision Electronics & US Army Research Lab's 'NVESD' MWIR-visible and LWIR-visible dataset by 18 and 35 % respectively. Our results show that this accounts for bridging the performance gap due to the modality difference by more than 40 %. To the best of our knowledge, this is the first attempt in using deep neural networks to bridge the modality gap in cross-modal face recognition. Figure 1 provides an overview of the approach.

This is the extended version of our previously published conference paper (Sarfraz and Stiefelhagen 2015) including a more thorough evaluation and analysis of the method. We will start by discussing some of the related work in Sect. 2.

Section 3 will detail the presented approach. We will conclude in Sect. 5 after presenting detailed experiments, results and implementation details in Sect. 4.

2 Related Work

One of the very first comparative studies on visible and thermal face recognition was performed by Socolinsky and Selinger (2002). They concluded that "LWIR thermal imagery of human faces is not only a valid biometric, but almost surely a superior one to comparable visible imagery." A good survey on single model and cross modal face recognition methods can be found in Zhou et al. (2014).

In the cross-modal (infra-red-visible) face recognition scenario, most of the earlier efforts focus only in the NIR to Visible matching. One of the first investigation by Yi et al. (2007) uses Linear discriminant analysis (LDA) and canonical correspondence analysis to perform linear regression between NIR and visible images. A number of approaches build on using local feature descriptors to represent the face. Liao et al. (2009) first used this approach on NIR to visible face recognition by processing face images with a difference of Gaussian (DoG) filter, and encoding them using multiblock local binary patterns. Gentle AdaBoost feature selection was used in conjunction with LDA to improve the recognition accuracy. Klare and Jain (2010) followed this work on NIR to visible face recognition by also incorporating SIFT feature descriptors and an LDA scheme. Lei and Li (2009) applied coupled spectral regression for NIR to visible recognition. Few methods have also focused on SWIR-to-visible face recognition (Ross and Hornak 2010; Nicolo and Schmid 2012). NIR or SWIR to visible face matching produces relatively better results as both the modalities are very similar because of the small spectral gap. Because of their limited use in the night-time surveillance applications, a much needed research focus is required in the thermal to visible matching domain.

Only recently some interest in the thermal to visible face matching has emerged. In the thermal domain, most of these

methods are evaluated in the MWIR to visible scenario. These methods employ similar techniques based on local features as in NIR to visible. A nice departure was proposed by Li et al. (2008) which is the only known method to perform recognition by trying to synthesize an equivalent visible image from the thermal counterpart. They provided some visual results and evaluated the method in a small hand collected data indicating an unsatisfactory performance. Another interesting local feature based approach has been put forward by Klare and Jain (2013), using local kernel prototypes. Their idea is to represent the two modalities by storing only similar vectors as prototypes during training using a similarity kernel space termed as prototype random subspace. They tested the method in different cross modal scenarios including MWIR to visible face recognition. Bourlai et al. (2012) proposed a MWIR-to-visible face recognition system based on simple preprocessing and feature matching. More recently Hu et al. (2014b) used a game theoretic based partitioning of the images to extract LBP type features and used SVM kernel matching for MWIR to visible face recognition achieving good results. A similar performance on the same dataset has been obtained in a very recent proposal (Hu et al. 2015) using partial least square 'PLS' based mapping between the MWIR and visible domain. MWIR sensors acquire higher resolution images than the LWIR. They, however, also operate at lower emissivity response to sense than the LWIR. The most difficult acquisition scenario is the usage of LWIR thermal images. Because of the long wave range sensing in LWIR, the images produced are usually of considerably lower resolution making it more challenging to match. On the other hand, pertaining to their higher emissivity sensing range they are also well suited to operate in total darkness. Attempts have been made to address face matching in the LWIR to visible domain. Choi et al. (2012) presented a PLS based regression framework to model the large modality difference in the PLS latent space. The most recent and best state-of-the-art results on the same dataset have been achieved by Hu et al. (2015). They used a discriminant PLS based approach by specifically building discriminant PLS gallery models for each subject by using thermal cross examples. They achieved a rank-1 identification of 50.4 % using two gallery images per subject and testing against all the probe LWIR thermal images. Chen and Ross (2015) put forward an interesting approach centred on learning common subspaces between visible and thermal facial images. The common subspaces are learnt via successive subspace learning process using factor analysis and common discriminant analysis on the image patches from both the domains. The projected vectors on these learnt subspaces are then directly matched. They provide convincing results on a private dataset. Finally, for facial feature detection in thermal images, Mostafa et al. (2013) presented a good method based on Haar features and Adaboost. They also performed face recognition experiments, but used both the visible and thermal modalities in the gallery and probe set.

Deep and large neural networks have exhibited impressive results for visible face recognition. Most notably, Facebook (Taigman et al. 2014) and Google (Schroff et al. 2015) showed excellent results using convolutional neural network to learn the face representation. Hu et al. (2014a) employ a feed forward network for metric-learning using pairs of same and different visible images. In contrast, we use a network to learn the non-linear regression projections between visible and thermal data without trying to learn any distance metric or image representation. A similar objective is employed in the domain adaptation based approaches e.g., Ganin and Lempitsky (2014), where the idea is to use the knowledge of one domain to aid the recognition in another image domain. Our model, while trying to learn a non-linear regression mapping, is simple and different from these in terms of the objective function and the way the network is used. We use a fully-connected feed forward network with the objective of directly learning a perceptual mapping between the two different image modalities.

3 Deep Perceptual Mapping (DPM)

The large perceptual gap between thermal and visible images is both because of the modality and resolution difference. The relationship between the two modalities is highly nonlinear and difficult to model. Previous approaches try to use non-linear function mapping e.g., using the kernel trick or directly estimating the manifold where the two could lie closer together. Since such a manifold learning depends on defining the projections based on a given function, they are highly data dependent and require a good approximation of the underlying distribution of the data. Such a requirement is hard to meet in practice, especially in the case of thermal–visible face recognition. Here, deep neural networks can learn, to some extent, the non-linear mapping by adjusting the projection coefficients in an iterative manner over the training set. As the projection coefficients are learned and not defined, it may be able to capture the data specific projections needed to bring the two modalities closer together. Based on this foresight we attempt to learn such projections by using a multilayer fully-connected feed forward neural network.

3.1 The DPM Model

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. Simple feed forward neural network provides a good fitting architecture for such an objective. We construct a deep network comprising $N + 1$

layers with $m^{(k)}$ units in the kth layer, where $k = 1, 2, \ldots, N$. 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 kth 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. Note, $h^{(0)}$ is the input. Similarly, the output of the most top level hidden layer can be computed as:

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

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. Since the objective of the model is to map one modality to the other, the final output layer of the network employs a linear mapping to the output in Eq. 1 to obtain the final mapped output:

$$\bar{x} = s\left(\mathbf{W}^{(N+1)}\mathbf{H}(x)\right) \qquad (2)$$

where $\mathbf{W}^{(N+1)} \in \mathbb{R}^{m^{(N)} \times d}$ and s is the linear mapping. Equation 2 provides the final perceptually mapped output $\bar{x}$ for an input x. 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.

$$arg \min_{W,b} \quad \mathbf{J} = \frac{1}{M}\sum_{i=1}^{M}(\bar{x}_i - t_i)^2 + \frac{\lambda}{N}\sum_{k=1}^{N}\left(\|\mathbf{W}^{(k)}\|_F^2 + \|\mathbf{b}^{(k)}\|_2^2\right) \qquad (3)$$

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 $\mathbf{X} = \{x_1, x_2, \ldots, x_M\}$ and $\mathbf{T} = \{t_1, t_2, \ldots, t_M\}$ from visible and thermal domains respectively, the objective of training is to minimize the function in Eq. 3 with respect to the parameters $\mathbf{W}$ and $\mathbf{b}$.

3.1.1 The DPM Training

There are some important design considerations for a meaningful training of the DPM network to provide an effective mapping from visual to thermal domain. First, the model is sensitive to the number of input and output units. A high dimensional input would require a very high amount of training data to be able to reasonably map between the two modalities. We propose to use the densely computed feature representations from overlapping small regions in the images. This proves very effective, as not only the model is able to capture the differing local region's perceptual differences well but also alleviate the need of large training images and nicely present the input in relatively small dimensions. The training set $\mathbf{X}$ and $\mathbf{T}$, then, comprises of these vectors coming from the corresponding patches from the images of the same identity. Note, only using the corresponding images of the same identity ensures that the model will only learn the differences present due to the modality as the other appearance parameters *e.g.* identity, expression, lighting etc., would most likely be the same. The network is trained, over the training set, by minimizing the loss in Eq. 3 w.r.t the parameters. Figure 1 encapsulates this process. By computing the gradient of the loss J in each iteration, the parameters are updated by standard back projection of error using stochastic gradient descent 'SGD' method of Glorot and Bengio (2010). For the non-linear activation function $g(z)$ in the hidden layers, we have used the hyperbolic tangent 'tanh' as it worked better than other functions e.g. sigmoid and ReLU on our problem. 'tanh' maps the input values between -1 to 1 using $g(z) = \frac{e^{2z}-1}{e^{2z}+1}$.

Before commencing training, the parameters are initialized according to a uniform random initialization. As suggested in Glorot and Bengio (2010), the bias $\mathbf{b}$ is initialized as 0 and the $\mathbf{W}$ for each layer is initialized using the uniform distribution $\mathbf{W}^{(k)} \sim U\left[-\frac{\sqrt{6}}{\sqrt{m^{(k)}+m^{(k-1)}}}, \frac{\sqrt{6}}{\sqrt{m^{(k)}+m^{(k-1)}}}\right]$, where $m^{(k)}$ is the dimensionality (number of units) of that layer.

3.2 Thermal–Visible Face Matching

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. Specifically, the mapped local descriptors from overlapping blocks of the visible gallery images are concatenated together to form a long vector. The resulting feature vector values are L_2-normalized and then matched with the similarly constructed vector directly from the probe thermal image. Note that, only visible gallery images are mapped through the learned projections using Eqs. 1 and 2 whereas the densely computed vectors from thermal images are directly concatenated into the final representation vector.

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. As we will show in the next section, without

any computational overhead the probes can be matched in real-time. Note, however, the presented DPM is independent of the direction of mapping. We observed only slight performance variations using the opposite thermal to visible mapping.

The identity of the probe image is simply determined by computing the similarity with each of the stored gallery image vectors and assigning the identity for which the similarity is maximum. Since the vectors are L_2-normalized, the cosine similarity can simply be computed by the *dot* product.

$$d(\bar{x}_i, t_j) = \bar{x}_i \cdot t_j \qquad \forall i = 1, 2, \ldots, G \qquad (4)$$

where t_j is the jth constructed probe thermal image vector and G is the total number of stored gallery vectors. Computationally, this is very efficient as at the test time, each incoming probe image can be compared with the whole gallery set $\bar{\mathbf{X}}$ by a single matrix-vector multiplication.

4 Experimental Results

In this section we present the evaluation of the proposed approach on the difficult thermal (LWIR & MWIR)-to-visible face recognition scenario. We report evaluations using typical identification and verification settings. The evaluation results, assessing the proposed DPM mapping, and the implementation details are discussed in the following subsections.

4.1 Database Description

Only few publicly available datasets include thermal and corresponding visible facial acquisitions. Among these, we use three very challenging datasets to evaluate the performance of the proposed cross-modal face matching method: UND Collection X1, Carl Database, and NVESD dataset. Each database was collected with different sensors containing image acquisitions under different experimental conditions e.g., different lighting, facial expressions, time-lapse, physical exercise and subject-to-camera range.

4.1.1 UND Collection X1

University of Notre Dame's UND collection X1 (Chen et al. 2005) is a challenging LWIR and visible facial corpus. The dataset was collected using a Merlin uncooled LWIR sensor and a high resolution visible color camera. The resolution of the visible images is 1600×1200 pixels and the acquired LWIR thermal images is 312×239 pixels. This depicts the challenge not only due to the large modality gap but also due to a large resolution difference. The dataset includes images in three experimental settings: expression (neutral, smiling, laughing), lighting change (FERET lighting and mugshot)

and time-lapse (subjects acquired in multiple sessions over time). Following the protocol used by the recently published competitive proposals on this dataset (Hu et al. 2015; Choi et al. 2012), we use all the available images in these three settings of the thermal probes to match against a single or multiple available high resolution visible gallery image/s. The dataset contains 4584 images of 82 subjects distributed evenly in visible and thermal domain. To compare our results, we use the exact same training and testing partitioning as used in the previous methods. The pool of subjects is divided into two equal subsets: visible and thermal images from 41 subjects are used for the gallery and probe images (partition A), while the remaining 41 subjects were used to train the DPM model. Note that the training and test sets are disjoint both in terms of images and subject's identities. All thermal images from partition-A are used as probes to compute recognition performance. The images are aligned based on provided eye and mouth locations and facial regions are cropped to 110×150 pixels. Figure 2 shows some sample visible and corresponding thermal images of one subject in the database.

4.1.2 Carl Database

The Carl dataset (Espinosa-Dur et al. 2013) is a relatively recently collected dataset containing images of 41 subjects acquired in visible, near-infrared and thermal (LWIR) spectrum. The dataset collects images under three illumination conditions (natural, artificial and infrared lighting) in 4 different sessions over time. The images primarily contain challenging appearance variations due to different illuminations, expressions, time-lapse and low image resolution. We use the thermal and visible corpus of the data which contains images across different lighting and acquisitions sessions totalling to $41 \times 5 \times 3 \times 4 \times 2 = 4920$ images, for 41 subjects, 3 illumination conditions (5 image/subject/illumination), 4-sessions in both visible and thermal domain. The resolution of the visible images is 640×480 and thermal (LWIR) images is 160×120 pixels. The segmented face images (output of a face detector) are provided by the authors. The face images, resized to 100×145 pixels, are not aligned w.r.t to eyes or nose position and therefore adds to the recognition challenge pertaining to misalignment errors. Since no earlier studies evaluate the full scale dataset in thermal to visible scenario, we provide the results to benchmark the full dataset in the following setting. We split the dataset equally with respect to the subjects. We use images of 20 subjects as the training set (1200 visible and 1200 thermal) and the rest as the probe. The training and test sets are, therefore, disjoint in terms of both subject's identities and images. Note, the probe set contains all the available thermal images of 21 subjects (from 3 illumination conditions in 4 sessions) totalling to 1260 images. The gallery set comprises visible images (1/subject and 2/subject

Fig. 2 Problem: Matching the thermal to the stored high resolution visible image. A wide modality gap exists between the visible and the thermal images of the same subject. Thermal images depict typical image variations, present, due to lighting, expressions and time-lapse

from session 1, natural illumination) of all the 41 subjects. This protocol will ensure an effective evaluation on the challenging thermal image variations due to illumination, time lapse and expressions against a large gallery.

4.1.3 NVESD Dataset

The NVESD dataset (Byrd 2013) was acquired as a joint effort between the Night Vision Electronic Sensors Directorate of the U.S. Army Communications-Electronics Research and the U.S. Army Research Laboratory (ARL). We use the thermal subset of the data containing images in both MWIR and LWIR spectrum. The NVESD dataset primarily examined two conditions in both spectrum: physical exercise (fast walk) and subject-to-camera range (1, 2, and 4 m). The dataset contains images of a total of 50 subjects, where images of 25 subjects are acquired in both before exercise and after exercise condition across all three ranges. The image resolution for the thermal sensors (MWIR and LWIR) is considerably higher at 640×480 pixels. The total number of thermal images in NVESD is 900 (450 MWIR, 450 LWIR) and the visible images are 450. The training set, correspond to subjects whose images are only acquired in before exercise condition, comprises 300 thermal images (150 MWIR, 150 LWIR) and 150 visible images. We train two separate models for MWIR-visible and LWIR-visible mapping using these training images.

Following the protocol in recently published state-of-the-art results on this dataset by Hu et al. (2015), we use the 25 subject subset (before and after exercise across all three ranges) as the probe/test set to examine the performance. The total number of probe thermal images is 600 (300 MWIR & 300 LWIR). The gallery set comprises the visible images of all the 50 subjects from before exercise condition at 1 m range. This protocol helps evaluate the performance in the presence of changed thermal signature (due to exercise) and subject-to-camera distance against the standard mugshot visible gallery image/s. The face images are normalised and aligned to 272×322 pixels as in Hu et al. (2015) in order to have a fair comparison.

4.2 Implementation Details

Here we provide the parameters used and related details for implementation. To compare the results to previous methods, we use a similar preprocessing for the thermal and visible images. We use the median filtering to remove the dead pixels in thermal images, then zero-mean normalization and afterwards used Difference of Gaussian 'DoG' filtering to enhance edges and to further reduce the effect of lighting, Figure 3 depicts aligned and preprocessed images of a subject. Dense features are then pooled from overlapping blocks in the image. The recognition performance varies between 2–5 % based on the descriptor used and the block size/overlap. A large overlap will produce more redundant information causing the network to saturate early while the block size effects the amount of structural information the network can still preserve while trying to map only the perceptual/modality difference. Based on our extensive evaluations we provide here the parameters that work well in practice. We have used a block size of 20×20 with a stride of 8 pixels at two-scales (Gaussian smoothed with $\sigma = 0.6$ and $\sigma = 1$) of the image. As many prior methods e.g. Klare and Jain (2013), Hu et al. (2014b, 2015), Choi et al. (2012) overwhelmingly stated the better performance of both the SIFT and HOG features in the thermal–visible face recognition. We experimented with both SIFT and HOG feature descriptors. However, here we present results with SIFT features as HOG produces 2–3 % inferior results than SIFT.

For the DPM model training, the number of units in the input/output and the hidden layers is important and varies the result. We use PCA to decorrelate and reduce the features to 64 dimensions. Each descriptor is further embedded with the block center position (x, y), measured from the image center, to have the spatial position intact, this helps the model to better learn the local region specific transformations in the hidden layers. The number of units in the input/output layers are, therefore, 66.

For the DPM network, we obtained better results with a minimum of 3-layers ($N = 3$ in Eqs. 1 and 2). Although, for comparison, we also report performance with using $N = 2$ i.e. one hidden layer. As our results show, the DPM mapping

Fig. 3 Aligned and preprocessed images of the corresponding visible and thermal images of a subject

is highly effective even with this shallow DPM configuration. The minimum number of units in the hidden layers, to ensure good results, is set empirically by using different combinations. Here, we report results using 200 units in each of the two hidden layers for the deep configuration. While in the case of shallow (1 hidden layer configuration), 1000 hidden units in the single layer provide close results. It is worth noticing, that the the results vary slightly about 1–4 % using different number of hidden units in the range 200–1000 in a layer. Finally the DPM is trained by pooling over 1 million patch vectors from the training set for each of the three datasets separately.

4.3 Results

We provide results of our evaluations using the settings described before. We first establish two baselines that would directly enable us to compare and appreciate the power of the proposed model. Along with these baselines we also compare the results of the proposed model with the previously published state of the art methods.

Baseline-1 As baseline-1 we use the same concatenated SIFT features without the DPM mapping. This would allow us to measure how much the learned mapping is contributing.

Baseline-2 As baseline-2 we use a state-of-the-art manifold learning method to embed the same SIFT vectors in a latent space using Partial Least Square (PLS) analysis. Vectors from both the thermal and visible images are projected through the learned PLS model in the latent space and then matched. The idea of PLS regression is to maximize the covariance between the dependent variable Y and a weighted sum of the

independent variable X by finding a weighing vector w. In our application, X and Y corresponds to the the visible and thermal vectors respectively. To obtain the weight matrix the PLS model is built by decomposing the $(n \times m)$ input matrix $X \subset \mathbb{R}^m$ and the response matrix $Y \subset \mathbb{R}^m$ into

$$
\begin{aligned}
X &= T P^T + E \\
Y &= U Q^T + F
\end{aligned}
\tag{5}
$$

The $(n \times p)$ matrices T and U are called scores that contain the latent vectors, the $(m \times p)$ matrices P and Q are called loadings and $(n \times m)$ matrices E and F are the residuals. By using a greedy algorithm, we can iteratively obtain a set of weight vectors stored in a matrix W. At the end of each iteration, the matrices X and Y are deflated by subtracting their rank-one approximations based on t and u until the desired number of latent vectors p is obtained. More details can be seen in Rosipal and Krämer (2006).

After obtaining W, we can compute the latent regression vectors from X to Y or vice versa by $Y = X B_v$ and $X = Y B_t$ where the regression model B is $B_v = W(P^T W)^{-1}$ and $B_t = W(Q^T W)^{-1}$ for the corresponding visible and thermal projections.

To ensure a proper baseline we perform PLS directly on the $m = 66$ dimensional SIFT vectors space. The number of latent vectors is found by cross validation on the training set and is set to $p = 20$ in all of our experiments. Each of the 66 dimensional SIFT vector from visible and thermal image is projected to this $p = 20$ dimensional latent space using the procedure mentioned above. All the projected vectors from an image are then concatenated and matched directly,

Table 1 Comparison of rank-1 identification accuracy (%) on UND X1

	Gallery size: # of visible images/subject		
	1/subject	2/subject	All available/subject
Baseline-1 (same features w/o DPM)	30.36	35.60	52.34
Baseline-2 (PLS on same features)	44.75	50.89	69.86
Choi et al. (2012)	–	49.9	–
Hu et al. (2015)	46.3	50.4	72.7
DPM-*1 hidden-layer* (ours)	50.67	58.26	79.57
DPM (ours)	55.36	60.83	83.73

Using all thermal images as probes and visible images in the gallery

similar to the procedure we adopt in baseline-1 and with the proposed DPM mapping.

Note that the previously published state of the art methods e.g., Hu et al. (2015) also uses a similar PLS framework. While we also compare with their method, the difference in our PLS-baseline is apparent not only in the choice of feature vectors but also the way PLS model is learned. Hu et al. (2015) use a discriminant PLS framework by building specifically discriminant PLS models for each identity in the database.

With these baselines one can appreciate the benefit of using the proposed DPM mapping by directly comparing the results on the same features using (1) direct matching and (2) using a state of the art regression mapping in a latent space obtained by PLS by exploiting the same amount of training data.

4.3.1 Identification Evaluation

Results on UND-X1 Table 1 presents the comparison results (Rank-1 identification score) with the baseline and the best published state-of-the-art results on the UND-X1 dataset. We present results using three different gallery settings. The most difficult and restricted setting with only one neutral visible image per subject in the gallery, using two visible images per subject in the gallery (the first two images, neutral and with smiling expression) and using all the available visible images per subject as gallery. Note that all the available thermal images/subject are used as probes. The number of images for each subject in the thermal probe set varies from a minimum of 4 to a maximum of 40 depicting appearance variations due to expressions, lighting and time-lapse. As the results show, we improve the state-of-the-art best published results of Hu et al. (2015) by 9 % in the 1-visible per subject gallery setting and more than 10 % in 2 per subject and all-available per subject cases. We also report the results using shallow DPM configuration (with 1-layer). Our results show that even this shallow DPM configuration well surpasses the best published results on this dataset.

Figure 4 presents the cumulative match characteristics 'CMC' curves to measure the rank-wise identification perfor-
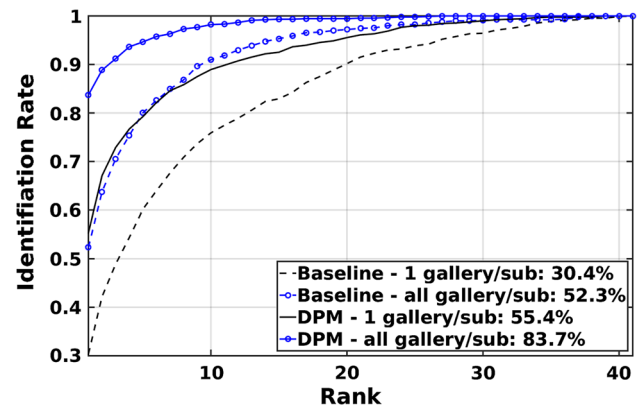

Fig. 4 Rank-wise score on UND X1: comparison of baseline and DPM performance with 1-visible image/sub and all available visible images/sub in the gallery

mance of the presented method. It shows the effectiveness of the proposed DPM mapping in comparison with the baseline-1 (same features without the DPM mapping).

Results on Carl Database The Carl dataset is the most challenging dataset because of the present strong facial appearance variations due to different illumination changes , time lapse, expressions, alignment errors and very low resolution LWIR images (160×120 pixels). To benchmark the full scale dataset, we present the identification results in a similar varying visible images/subject gallery setting. As mentioned earlier, the probe set contains all the thermal (LWIR) images from 21 test subjects and matched against a gallery containing visible images of all the 41 subjects. Table 2 presents the obtained Rank-1 scores of both the baseline and the presented DPM mapping across different gallery image settings. Note while the Rank-1 scores on Carl dataset, despite a more difficult imaging scenario than UND-X1, are comparable in the 1 image per subject and 2 image per subject gallery setting. The relatively low Rank-1 score when using all available visible images per subject in the gallery is understandable considering the large number of gallery images (60 visible images/subject = 2460 images form 41 subjects), where these visible images contain very difficult illumination variations (natural light, artificial light and infra-red illumination). We

Table 2 Rank-1 identification accuracy (%) on Carl database

| | Gallery size: # of visible images/subject | | |
	1/subject	2/subject	all available/subject
Baseline-1 (same features w/o DPM)	26.50	33.83	53.08
Baseline-2 (PLS on same features)	31.75	34.66	51.58
DPM-*1 hidden-layer* (ours)	50.42	59.92	69.33
DPM (ours)	56.33	60.08	71.00

Using all thermal images as probes and visible images in the gallery
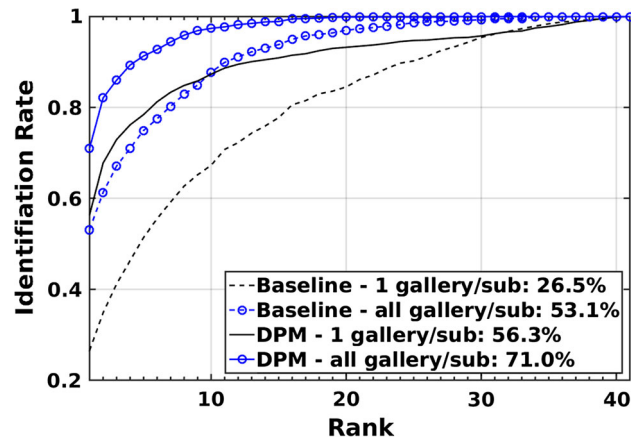
Fig. 5 Rank-wise score on Carl Dataset: comparison of baseline and DPM performance with 1-visible image/sub and all available visible images/sub in the gallery

Table 3 Rank-1 identification accuracy (%) on NVESD

| | Subject-to-camera range: | | |
	1 m	2 m	4 m
(a) MWIR-visible			
Before exercise	(92)**98**	(70)**86**	(58)**82**
After exercise	(88)**100**	(68)**92**	(48)**74**
(b) LWIR-visible			
Before exercise	(70)**94**	(64)**92**	(30)**84**
After exercise	(64)**96**	(62)**98**	(28)**64**

Using Before Exercise and After Exercise thermal probe images at each of the three distance ranges. (a) Using MWIR probe images, (b) using LWIR probe images. Gallery includes 2 visible image/subject acquired in before exercise condition at 1 m range of all the 50 subjects. Parenthesis (·) depict accuracies of the current best published results of Hu et al. (2015), Our results are depicted in bold

also tested the effect of overall gallery size i.e., only having visible images of the 21 test subjects in the gallery. Here, the Rank-1 scores using DPM improves as expected. We obtained an accuracy of 66.58 % with the 1-image per subject, 72.92 % with 2-images per subject and 82.50 % with all available images per subject gallery setting.

Figure 5 presents CMC curves to measure the rank-wise identification performance of the presented method and the baseline-1 on Carl dataset.

Results on NVESD Dataset The NVESD dataset helps evaluate the method on both MWIR and LWIR thermal domains. Given the image acquisitions from relatively better thermal sensors (640×480 pixel thermal images), the evaluations on NVESD resulted in much better performance. We evaluate and compare the Rank-1 identification results with that of the state-of-the-art results of Hu et al. (2015), using the same test protocol.

We primarily evaluate the effect of exercise on the thermal signatures of the subjects across different subject-camera ranges (1, 2, and 4 m). For this, the 25 subject subset both from MWIR and LWIR thermal sets is used as probe and matched against 2-visible images per subject in the gallery. The two visible-image per subject gallery images are the neutral mugshots, acquired at 1m range in before exercise condition. Following Hu et al. (2015), we prepare 6 probe

sets for each MWIR and LWIR thermal test sets (before exercise and after exercise condition $\times$ three subject-to-camera ranges). Each probe set contains 50 images (2/subject) in each of the specific condition-range pair test. Table 3 presents the Rank-1 scores for MWIR-Visible and LWIR-Visible scenario of this experiment. As the results depict, DPM mapping improves the state-of-the-art by a considerable margin, especially at longer ranges. An intriguing result is the better accuracy at 1m and 2m range of after exercise condition than before exercise. A closer look at the misclassified image/s of before-exercise probe sets at these ranges reveals that the effect seems to be with facial expressions in these pairs causing the similarity scores falling slightly below the rank-1 threshold. Given that the training has not seen the images in after-exercise condition, this also implies that the DPM mapping is not effected by the small thermal change present in the images due to this.

For MWIR-visible scenario we obtained on average a Rank-1 score of 88.6 % versus an average of 70.6 % of Hu et al. (2015). We, therefore, improve the previous results by 18 % on the NVESD's MWIR-visible dataset. Similarly for LWIR-visible, we obtained on average Rank-1 score of 88 % versus an average of 53 % of Hu et al. (2015). Thus, on the NVESD's LWIR-visible dataset, we improve the result by 35 %. These results show that the learned DPM mapping is relatively stable and effective against the thermal signature

Table 4
MWIR-visible—overall rank-1 identification accuracy (%) on NVESD

	Gallery size: # of visible images/subject		
	1/subject	2/subject	All available/subject
Baseline-1 (same features w/o DPM)	72.00	75.66	92.00
Baseline-2 (PLS on same features)	77.66	79.33	97.66
DPM-*1 hidden-layer* (ours)	**86.66**	87.33	97.66
DPM (ours)	86.00	**87.66**	**98.66**

Bold values depict the best scores obtained in that column for the presented method
Using all MWIR images as probes and visible images in the gallery

Table 5 LWIR-visible—overall rank-1 identification accuracy (%) on NVESD

	Gallery size: # of visible images/subject		
	1/subject	2/subject	All available/subject
Baseline-1 (same features w/o DPM)	61.00	64.33	88.00
Baseline-2 (PLS on same features)	66.66	67.00	93.66
DPM-*1 hidden-layer* (ours)	**83.67**	82.00	96.67
DPM (ours)	82.67	**85.00**	**97.33**

Bold values depict the best scores obtained in that column for the presented method
Using all LWIR images as probes and visible images in the gallery
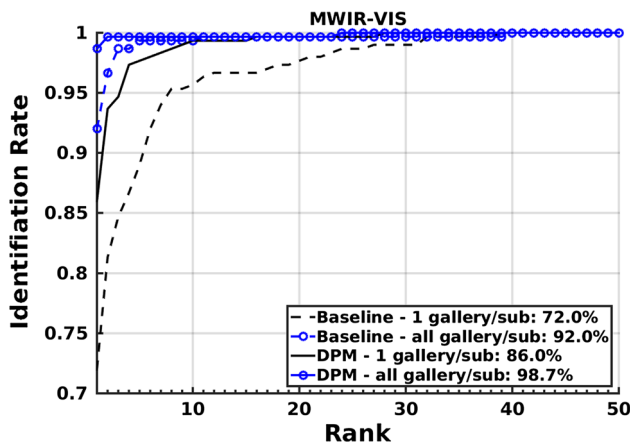
Fig. 6 Rank-wise score on NVESD MWIR-visible dataset: comparison of baseline and DPM performance with 1-visible image/sub and all available visible images/sub in the gallery
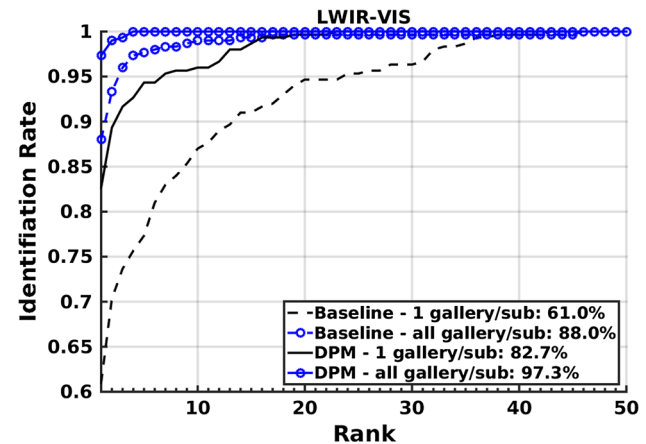
Fig. 7 Rank-wise score on NVESD LWIR-visible dataset: comparison of baseline and DPM performance with 1-visible image/sub and all available visible images/sub in the gallery

change due to exercise while performs significantly better on image acquisition at longer range. Tables 4 and 5 provide the overall Rank-1 accuracy of the Baselines and the DPM in different visible-image/subject in the gallery settings on MWIR-visible and LWIR-visbile NVESD datasets.

Figures 6 and 7 provide the similar rank-wise performance of our method and compares it with the baseline-1 features in the 1-visible per subject and the all-available visible per subject gallery setting.

4.3.2 Verification Evaluation

We also evaluate the verification accuracy of such cross-modal scenario. We report here the results by using 2 visible images/subject in the gallery on UND-X1, Carl and NVESD datasets. Given the size of our thermal probe set, this amounts

to having 1792 genuine and 71680 imposter attempts on UND-X1, 2520 genuine and 100800 imposter attempts on Carl database and 600 genuine and 29400 imposter attempts on both NVESD's LWIR and MWIR datasets. Figures 8, 9, 10 and 11 present the ROC curves measuring the verification performance gain over the baseline-1 using the DPM mapping.

4.3.3 Effect of Modality Gap

Finally, we present the experiment to measure the effect of the modality gap. Keeping everything fixed i.e., using the same baseline-1 features and settings, we compute the Rank-1 identification score within the same modality. We use one-thermal image per subject (the neutral frontal image) to form the thermal gallery and test against all the remaining thermal
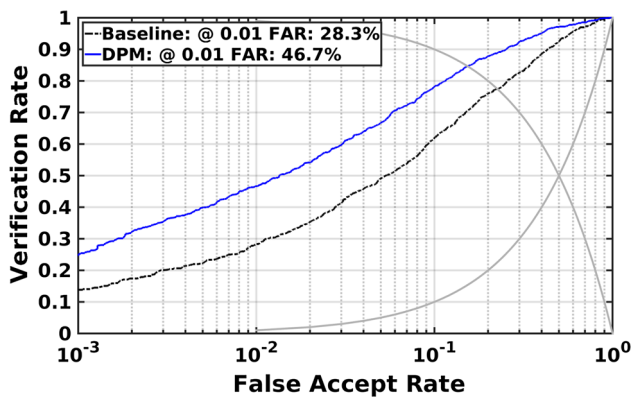
Fig. 8 Verification performance on UND X1: all thermal probes with 2 visible image/sub in gallery
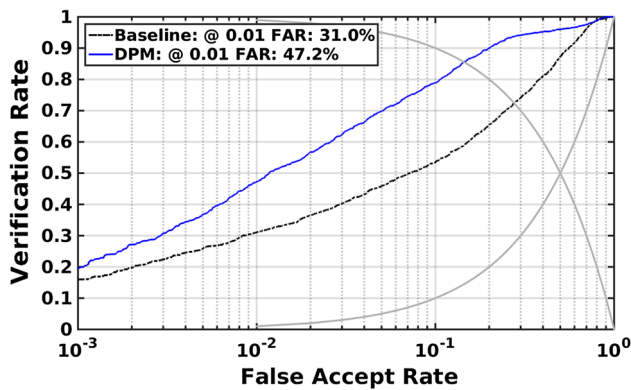

Fig. 9 Verification performance on Carl Database: all thermal probes with 2 visible image/sub in gallery
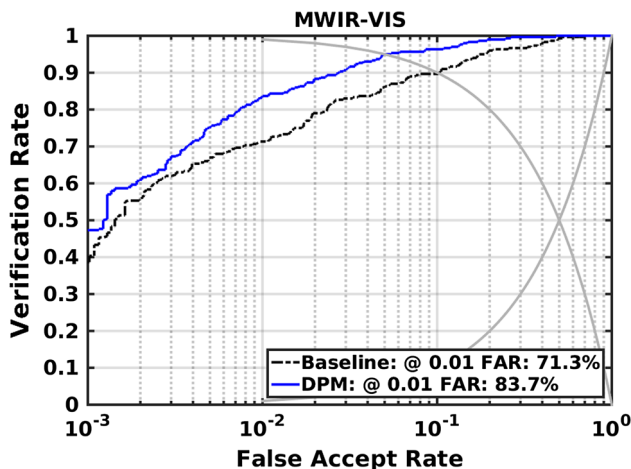

Fig. 10 Verification performance on NVESD MWIR dataset: all thermal probes with 2 visible image/sub in gallery

images as probe. This is the same setting as we have used in the cross-modal thermal–visible case. Table 6 compares the results of Thermal–Thermal and Thermal–Visible identification on all three datasets to quantify the effect of modality gap.
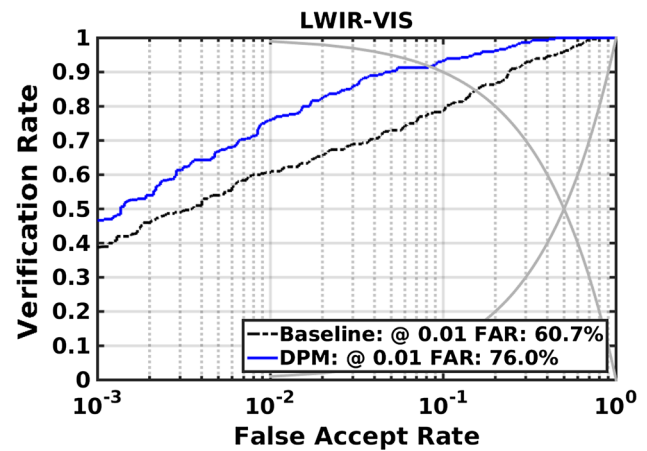

Fig. 11 Verification performance on NVESD LWIR dataset: all thermal probes with 2 visible image/sub in gallery

On UND-X1, we obtain a Rank-1 score of 89.7 % in the Thermal–Thermal identification scenario. While the rank-1 identification in the corresponding Thermal–Visible scenario (using the same baseline-1 features) is 30.3 %. This amounts to 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. 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 % on UND-X1. Similar observations can be made by interpreting the results on other datasets as given in Table 6.

4.3.4 Computational Time

Training the DPM on 12 cores 3.2-GHz CPU takes between 1–1.5 h on MATLAB. Preprocessing, features extraction and mapping using DPM only takes 45 ms for one image. This is even less in the testing case since no mapping is required for thermal images. At test time identifying one probe only takes 35 ms. Since we are using just the dot product between the extracted probe vector and the gallery set, this is therefore very fast and capable of running in real-time at ∼28 fps.

5 Discussion and Conclusions

While providing quantitative evidence of the effectiveness of the proposed approach, we can also compare our results qualitatively with some other methods that evaluate the thermal–visible face recognition on similar private datasets. (Bourlai et al. 2012) reported results on a MWIR-visible face recognition task using a 39 subject dataset with four visible images per subject and three MWIR images per subject as probes with 1024 × 1024 pixel resolution. They reported a

Table 6 Performance drop due to modality gap: Rank-1 identification using 1 image/subject as gallery in thermal–thermal and thermal–visible matching using baseline-1 features

	Effect of Modality gap: Performance with 1 Gallery image/subject			
	Therm-Therm	Therm-Vis	Therm-Vis(DPM)	Modal-gap bridged
UND-X1	89.4	30.3	55.3	∼ 42 %
Carl Dataset	61.7	26.5	56.3	∼ 84 %
NVESD MWIR	98.6	72	86	∼ 48 %
NVESD LWIR	97	61	82.6	∼ 40 %

rank-1 performance of 53.9 %. Our result of 86 % on the NVESD MWIR-visible dataset with just one visible image per subject in the gallery can be qualitatively compared to these results. Similarly, Chen and Ross (2015) evaluated their method on a subset of the Carl dataset with 5 visible image per subject in the gallery and 5 thermal image per subject as probe. They, however, evaluate in a cross dataset scenario *i.e.*, subspaces learned on another similar private dataset and tested on the Carl data subset. They reported a best rank-1 performance of 75.6 %. Compared to this, we evaluate the full scale Carl dataset with 60 thermal images per subject in the probe set. Our results with just 2 visible image per subject and multiple visible images per subject in the gallery, as depicted in Table 2, provide a qualitative motivation for the effectiveness of the proposed method. Finally, in a very recent proposal (Riggan et al. 2016) the authors evaluate their own implementation of our presented DPM method on polarimetric thermal to visible face recognition problem achieving one of the highest performance on this dataset. This further substantiates the applicability of the presented method on a general cross modal image matching scenario.

Given the very high recognition ability of Convolutional Neural Network (CNN) features in the conventional visible face recognition, a possible future direction to further improve these results is to directly train a CNN model. On the cross-modal face matching problem, the CNN model can be trained in a similar metric learning fashion as presented by Google's FaceNet (Schroff et al. 2015) or Oxford's VGG deep face model (Parkhi et al. 2015). The current limitation for a meaningful training of such a model is the lack of enough thermal facial images. It is hoped that, provided, the vast and demanding application of cross-modal face matching and availability of low-priced accessible thermal sensors, more and more thermal facial data can be acquired. Note, a direct application of the current pre-trained CNN models on the thermal images do not yield meaningful results. These models are trained on three channel RGB colour images where as the thermal images are single channel 16-bit images. Our initial investigations to get a baseline using the pre-trained CNN models of (Schroff et al. 2015; Parkhi et al. 2015), by just replicating the thermal image in three channels, did not provide acceptable results. Here, however, a similar perceptual mapping strategy may be further investigated by training

a CNN using only the image patches from both domains. Such a direction may provide interesting results even with relatively less training data.

Conclusively, thermal–visible face recognition is a very difficult problem due to the inherent large modality difference. Our presented method is very effective and has benefits for many related applied computer vision and domain adaptation problems from image matching, detection to recognition. Similarly it is also very attractive for remote sensing applications, where it is a common problem to register and match images from different modalities (e.g., coming from different satellites). This can be used as a very effective prior step to bridge the modality and/or large resolution difference before using the representation vectors in any common learning scheme. The presented DPM approach is very useful, easy to train, and real-time capable due to little computational overhead and it provides a practical solution for a large related application industry.

References

Bourlai, T., Ross, A., Chen, C., & Hornak, L. (2012). A study on using mid-wave infrared images for face recognition. In *SPIE Defense, Security, and Sensing, International Society for Optics and Photonics* (pp 83,711K–83,711K).

Byrd, K. (2013). Preview of the newly acquired nvesd-arl multimodal face database. In: *Proceedings of SPIE*, vol *8734* (pp. 8734–8734).

Chen, C., & Ross, A. (2015). Matching thermal to visible face images using hidden factor analysis in a cascaded subspace learning framework. *Pattern Recognition Letters*. doi:10.1016/j.patrec.2015.06.021. http://www.sciencedirect.com/science/article/pii/S0167865515001932.

Chen, X., Flynn, P. J., & Bowyer, K. W. (2005). IR and visible light face recognition. *Computer Vision and Image Understanding*, *99*(3), 332–358.

Choi, J., Hu, S., Young, S. S., & Davis, L. S. (2012). Thermal to visible face recognition. In *SPIE Defense, Security, and Sensing, International Society for Optics and Photonics* (pp. 83,711L–83,711L).

Espinosa-Dur, V., Faundez-Zanuy, M., & Mekyska, J. (2013). A new face database simultaneously acquired in visible, near-infrared and thermal spectrums. *Cognitive Computation*, *5*(1), 119–135. doi:10.1007/s12559-012-9163-2.

Ganin, Y., & Lempitsky, V. (2014). Unsupervised domain adaptation by backpropagation. arXiv:1409.7495.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics* (pp 249–256).

Hu, J., Lu, J., & Tan, Y.P. (2014a). Discriminative deep metric learning for face verification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1875–1882).

Hu, S., Gurram, P., Kwon, H., & Chan, A. L. (2014b). Thermal-to-visible face recognition using multiple kernel learning. In *SPIE Defense+ Security, International Society for Optics and Photonics* (pp 909,110–909,110).

Hu, S., Choi, J., Chan, A. L., & Schwartz, W. R. (2015). Thermal-to-visible face recognition using partial least squares. *Journal of the Optical Society of America*, *32*(3), 431–442.

Klare, B., & Jain, A. (2010). Heterogeneous face recognition: Matching nir to visible light images. In *20th International Conference on Pattern Recognition (ICPR), 2010* (pp. 1513–1516).

Klare, B. F., & Jain, A. K. (2013). Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(6), 1410–1422.

Lei, Z., & Li, S. Z. (2009). Coupled spectral regression for matching heterogeneous faces. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (pp 1123–1128).

Li, J., Hao, P., Zhang, C., & Dou, M. (2008). Hallucinating faces from thermal infrared images. In *15th IEEE International Conference on Image Processing, 2008. ICIP 2008* (pp. 465–468).

Li, S. Z., Chu, R., Liao, S., & Zhang, L. (2007). Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(4), 627–639.

Liao, S., Yi, D., Lei, Z., Qin, R., & Li, S. Z. (2009). Heterogeneous face recognition from local structures of normalized appearance. In: *Advances in Biometrics*, Springer (pp. 209–218).

Mostafa, E., Hammoud, R., Ali, A., & Farag, A. (2013). Face recognition in low resolution thermal images. *Computer Vision and Image Understanding*, *117*(12), 1689–1694.

Nicolo, F., & Schmid, N. A. (2012). Long range cross-spectral face recognition: Matching swir against visible light images. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1717–1726.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

Riggan, B. S., Nathaniel, J. S., & Shuowen, H. (2016). Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Rosipal, R., & Krämer, N. (2006). Overview and recent advances in partial least squares. In *Subspace, latent structure and feature selection*, Springer (pp. 34–51).

Ross, T. B. N. K. A., & Hornak, B. C. L. (2010). Cross-spectral face verification in the short wave infrared (swir) band.

Sarfraz, M. S., & Stiefelhagen, R. (2015). Deep perceptual mapping for thermal to visible face recognition. In *British Machine Vision Conference*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. arXiv:1503.03832.

Socolinsky, D. A., & Selinger, A. (2002). A comparative analysis of face recognition performance with visible and thermal infrared imagery. Tech. rep., DTIC Document.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014* (pp. 1701–1708).

Yi, D., Liu, R., Chu, R., Lei, Z., & Li, S. Z. (2007). Face matching between near infrared and visible light images. In: *Advances in Biometrics*, Springer (pp 523–530).

Zhou, H., Mian, A., Wei, L., Creighton, D., Hossny, M., & Nahavandi, S. (2014). Recent advances on singlemodal and multimodal face recognition: A survey. *IEEE Transactions on Human Machine Systems*, *44*(6), 701–716.