A Parametric Study of Deep Perceptual Model on Visible to Thermal Face Recognition

Wei-Ta Chu
Department of Computer Science and Information Engineering
Advanced Institute of Manufacturing with High-tech Innovations
National Chung Cheng University, Taiwan
Email: wtchu@ccu.edu.tw

Jo-Ning Wu
National Chung Cheng University, Taiwan
Email: dylane1024@gmail.com

***Abstract*—Recently deep perceptual mapping (DPM) based on auto-encoder provides the state-the-art thermal to visible face recognition. Features extracted from patches of a long-wave infra-red (LWIR) face image are transformed into a space by an auto-encoder, such that features from infra-red images are comparable with features from visible images. In this paper, we comprehensively evaluate DPM with different settings, in order to build a reference study for future research.**

I. Introduction

Lighting/illumination variations cause significant visual appearance changes and largely degrade face recognition performance. Some studies, therefore, have been proposed to focus on recognizing infra-red face images. Thermal signatures emitted by skin tissues are acquired by passive thermal sensors, and thus infra-red faces are invariant to lighting. Such characteristics makes infra-red face recognition quite potential in night-time surveillance applications or access control systems with the requirement of privacy protection.

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

Infra-red images can be categorized according to the wavelengths sensed, including near infra-red 'NIR' (0.74 μm – 1 μm), short-wave infra-red 'SWIR' (1 μm – 3 μm), mid-wave infra-red 'MWIR' (3 μm – 5 μm), and long-wave infra-red 'LWIR' (8 μm – 14 μm). NIR and SWIR are light-reflection-based and their visual appearance are similar to visible images. Most prior studies focused on NIR or SWIR, and promising recognition performance has been achieved [1][2][3][4][5]. On the contrary, MWIR and LWIR images are captured depending on material emissivity and temperature, and give rise to much severe challenges because of the significant gap between the visible spectrum and infra-red spectrum. Currently only few studies have focused on MWIR and LWIR face images, and still only limited performance can be achieved [6][7].

Among the current studies on LWIR face recognition, the deep perceptual mapping (DPM) method [6] acts as the state of the art. The relationship between visible spectrum and infra-red spectrum is highly nonlinear. Previous works discovered this nonlinear mapping by manifold learning or kernel mapping, but nonlinearity of realistic data was not well described. Inspired by the success of deep learning methods, the DPM method learns such mapping based on an auto-encoder. SIFT descriptors are extracted from overlapping patches of visible face images, and then, by the leant auto-encoder, are projected into a space commonly shared with the features extracted from infra-red face images. Given a visible image, its identity is determined by finding the infra-red image that has the feature vector most similar to the transformed feature vector coming from the visible image[1]. The DPM method was the first one to bring the deep learning method to thermal face recognition, and has yielded significant improvement over previous works.

The DPM method finds nonlinear mapping between patches from infra-red images and visible images. In this paper, we will study performance variations of DPMs with different settings based on the thermal face dataset collected by Nagoya University [7] (named as the NU dataset in the following). 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, respectively. Therefore, the visible face and the thermal face of the same individual are well aligned. We think such alignment is important for us to clearly study performance variations of DPM with different parameters.

II. Deep Perceptual Mapping

Our work is based on one of the state-of-the-art thermal-to-visible recognition methods, i.e., deep perceptual mapping [6]. DPM is designed to map features extracted from one modality to another modality by an auto-encoder consisting of $N+1$ layers. Given $\boldsymbol{x} \in \mathbb{R}^d$, each layer projects the input by a learned projection matrix $\boldsymbol{W}$ and the nonlinear activation function $g(\cdot)$. The output of the kth layer is $\boldsymbol{h}^{(k)} = g(\boldsymbol{W}^{(k)}\boldsymbol{h}^{(k-1)} + \boldsymbol{b}^{(k)})$, where $\boldsymbol{b}^{(k)}$ is a bias vector. The Nth layer is a linear mapping to make prediction from hidden layers, and forms a vector of targeted dimensionality, i.e, $\hat{\boldsymbol{x}} =$

[1]In [6], the query is a visible face image, and features extracted from it are transformed to match with features extracted from infra-red images. We actually can do the reverse, i.e., taking infra-red images as the query and transforming features to match with that extracted from visible images. According to our experiments, performance of these two schemes is similar.

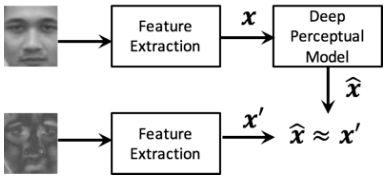

Fig. 1. An illustration of the idea of the deep perceptual model.

(a) thermal image (b) visible image

Fig. 2. Sample images of the NU database.

$\boldsymbol{W}^{(N)}\boldsymbol{h}^{(N-1)}$. The initial $\boldsymbol{h}^{(0)}$ is the input $\boldsymbol{x}$. To determine the projection matrices $\boldsymbol{W} = \{\boldsymbol{W}^{(0)}, \boldsymbol{W}^{(1)}, ..., \boldsymbol{W}^{(N)}\}$ and bias vectors $\boldsymbol{b} = \{\boldsymbol{b}^{(0)}, \boldsymbol{b}^{(1)}, ..., \boldsymbol{b}^{(N)}\}$, the mean square error between the features transformed by the auto-encoder, and the features extracted from another modality, is calculated as the objective function.

Fig. 1 illustrates the idea of the DPM. From the visible face image and its corresponding thermal face image, we extract features as $\boldsymbol{x}$ and $\boldsymbol{x}'$, respectively. The DPM is constructed to transform $\boldsymbol{x}$ into $\hat{\boldsymbol{x}}$ such that the transformed vector $\hat{\boldsymbol{x}}$ is similar to the feature $\boldsymbol{x}'$ extracted from the thermal face image.

The model mentioned above provides a good foundation to find appropriate nonlinear mapping between two modalities, yet data processing also plays a very important role in achieving promising performance. In [6], densely computed feature representations from overlapping regions in the images are used as the input vector $\boldsymbol{x}$'s. Particularly SIFT descriptors are extracted from 20×20 image patches with a stride of 8 pixels, in the images of 110×150 pixels. Each image patch is represented by as a SIFT descriptor, with dimensionality reduced to 64 with principal component analysis, associated with the patch center position (x, y) to embed spatial information, yielding a 66-dimensional vector. With such local representation, local perceptual difference can be described, and the requirement of large training images is mitigated because we can pool a large number of patches from the limited set of training images. According to [6], different patch sizes or overlapping cause 2–5% performance variations, and SIFT outperforms HOG with 3% performance improvement.

In [6], the suggested DPM framework contains three layers, with each layer consisting of 200 units. Given the feature sets coming from visible training patches $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\}$ and thermal training patches $\boldsymbol{T} = \{\boldsymbol{t}_1, ..., \boldsymbol{t}_M\}$, the network parameters are determined by the stochastic gradient descent (SGD) method, with the hyperbolic tangent "tanh" as the nonlinear activation function $g(\cdot)$. Note that these training patches are in pairs and with identity and spatial correspondence. For example, the visible image where $\boldsymbol{x}_i$ comes from the same individual as the thermal image where $\boldsymbol{t}_i$ comes from. In addition, the location of the patch where $\boldsymbol{x}_i$ comes from the same patch as where $\boldsymbol{t}_i$ comes from. When testing, given the probe image represented as a collection of patches $\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_K\}$ the transformed vectors $\{\hat{\boldsymbol{x}}_1, ..., \hat{\boldsymbol{x}}_K\}$ are concatenated as a vector, which is compared with the concatenated vectors of each gallery image based on the cosine similarity. The identity of probe image is determined as the identity of the gallery image with the maximum cosine similarity value to it.

III. Evaluation Protocol

A. Database

In order to conduct a systematic and convincing parametric study, we need a dataset where well-aligned visible face images and thermal face images are available. In other words, thermal and visible face images of the same individual were captured in very similar pose, expression, and other visual conditions. By avoiding the fluctuation of visual conditions, we can focus on evaluating how different parameters influence recognition performance.

In this work, we choose to utilize the Nagoya University (NU) database [7] for our study. The NU database contains 180 Japanese people (169 males and 11 females), where five pars of thermal images and visible image were captured for each individual. The ordinary camera capturing visible spectrum 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 dataset like the USTC-NVIE database [10], 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 8–14 μm. The corresponding thermal and visible images were captured at the same time, and were underwent the same preprocessing before used. After cropping, calibration, and resizing, resolution of both types of images is 56×64 pixels. Fig. 2(a) and Fig. 2(b) show a sample image pair from the NU database.

B. Parametric Study

We attempt to verify effectiveness of the DPM framework based on the NU database. In the evaluation protocol of [7], 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 to constructed the DPM. 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.

We implement the DPM framework consisting of four fully-connected layers. Outputs of the first three layers are

TABLE I
DETAILED CONFIGURATION OF THE IMPLEMENTED DPM FRAMEWORK.

input (56×56 thermal images)			
fully-connected (2048 nodes) activation fun.: ReLU dropout(0.1) Adding Gaussian noise	fully-connected (2048 nodes) activation fun.: ReLU dropout(0.1) Adding Gaussian noise	fully-connected (2048 nodes) activation fun.: ReLU dropout(0.1) Adding Gaussian noise	fully-connected (2048 nodes) activation fun.: ReLU

intentionally added with Gaussian noise to mitigate overfitting. The activation function of each layer is ReLU, the objective function is the mean square error between transformed vector and ground truth vector, and the optimization algorithm is Adam. The training process was conducted in 40 epochs, with mini-batch size 200. Table I shows detailed configuration of the implemented DPM framework.

Based on the NU dataset, we would like to investigate how the following factors yield performance variations.

- Features: The DPM method shown in [6] extracts SIFT features from patches and then transforms with an auto-encoder structure. They briefly said that SIFT features outperforms HOG by 3%. In our work, we carefully evaluate SIFT only, HOG only, and the combination of them, and show performance comparison.
- Patch size: In [6], the size of patches is set as 20×20 pixels to extract features from 110×150 face images. Although they briefly mentioned that performance varies between 2–5% when different sizes of patches are used, detailed evaluation was missing. In our work, we will evaluate patches of 4×4 pixels, 8×8 pixels, 10×10 pixels, and 16×16 pixels on the face images of 55×75 pixels in the NU database. The strides for these four settings are 2 pixels, 4 pixels, 4 pixels, and 8 pixels, respectively.
- Noise level: Noise level is the special design of [7]. For testing thermal images of 10 people, other than these 10 people's visible faces, visible faces of 20 other people never seen at the training stage are put into the candidate pool. In our work, we would like to investigate how the noise level influences recognition performance.
- Glasses: In the NU database, among the 180 individuals, 21 individuals are wearing glasses. For these individuals, five visible/thermal images were captured without glasses, and five other visible/thermal images were captured with glasses. In most of the experiments reported in this paper, we only consider images without glasses for training and testing. However, we will also conduct an experiment studying the influence of glasses.

IV. EVALUATION RESULTS

A. Performance Variations given by Different Features

We first evaluate the performance variations yielded by different features and compare the DPM method with [7]. As can be seen in Table II, no matter which features are used, the DPM method largely outperforms [7]. This confirms the superiority of deep-based methods, especially DPM, on thermal face recognition. In [7], the nonlinear mapping between features extracted from two different modalities is determined by canonical correlation analysis (CCA). By comparing HOG with SIFT, we confirm that SIFT outperforms HOG. In [6], they said that SIFT features outperforms HOG by 3%, and SIFT features outperforms HOG by 16% based on the NU database. We are also interested in if the performance can be further improved if we combine them. However, as shown in the last column of Table II, combining them slightly decreases performance. In the following experiments, SIFT features are used as the image descriptor in the DPM framework.

TABLE II
AVERAGE RECOGNITION ACCURACIES OF [7] AND THE DPM FRAMEWORK WITH DIFFERENT FEATURES, BASED ON THE NU DATABASE.

	[7]	HOG	SIFT	SIFT+HOG
Avg. Accuracy (%)	23.13	43.50	**59.50**	57.13

TABLE III
AVERAGE RECOGNITION ACCURACIES OF [7] AND THE DPM FRAMEWORK WITH DIFFERENT PATCH SIZES, BASED ON THE NU DATABASE AND THE NOISE LEVEL MENTIONED IN [7].

	4×4	8×8	10×10	16×16
Avg. Accuracy (%)	53.88	58.50	**59.50**	53.63

B. Performance Variations given by Different Patch Sizes

We then evaluate the influence of different patch sizes on recognition performance. Table III shows the average recognition accuracies of the DPM framework with different patch sizes, based on the NU database. Comparing performance yielded by DPMs with different patch sizes, we see that the best performance can be obtained based on patches of 10×10 pixels. In the following experiments, patches of 10×10 pixels, with stride 4 pixels, are used.

C. Performance Variations given by Different Noise Levels

To mimic the realistic scenario, a training and testing protocol was set in [7]. We name its setting as (tr=150, ts=10, noise=20), indicating that data of 150 people are used for training, and the other 10 people are for testing. In addition

TABLE IV
AVERAGE RECOGNITION ACCURACIES OF THE DPM FRAMEWORK WITH DIFFERENT NOISE LEVELS, BASED ON THE NU DATABASE.

	(tr=150, ts=10, noise=0)	(tr=150, ts=10, noise=10)
Avg. Acc.	62.75	62.50
	(tr=150, ts=10, noise=20)	(tr=140, ts=10, noise=30)
Avg. Acc.	**59.50**	46.80
	(tr=130, ts=10, noise=40)	
Avg. Acc.	40.29	

TABLE V
AVERAGE RECOGNITION ACCURACIES OBTAINED BY THE DPM FRAMEWORK WITH THREE DIFFERENT TRAINING/TESTING SCHEMES.

	(1)	(2)	(3)
Avg. Accuracy (%)	59.50	57.52	**62.51**

to the 10 people for testing, 20 other people who were never seen by the model are added to the candidate pool to make the recognition task more complex. We are interested in how different noise levels yield performance variations.

Table IV shows performance variations yielded by different noise levels. From the first column to the third column, we see that, with the same number of training data, performance slightly decreases as the noise level increases. This is expectable, and the trend of slight decreasing shows the robustness of the DPM method. From the fourth column and the fifth column, we increase noise but decrease the number of training data. In these cases, the average accuracy significantly decreases.

D. Performance Variations with or without Glasses

We evaluate three training/testing schemes to show the influence of wearing glasses. (1) Both the training data and the testing data do not include images with glasses. (2) The training data do not include images with glasses, but the testing data include images with glasses. (3) Both the training data and the testing data include images with glasses. Table V shows average recognition accuracies with these three schemes. Comparing with the first two schemes, it is not surprising that performance degrades when the testing data include images with glasses. Interestingly, by seeing the third column, we see that the best performance can be obtained when both training data and testing data include images with glasses. Intensity contrast of thermal face images is relatively weaker, and thus weaker SIFT descriptors can be extracted. In this case, glasses usually present clear corners or boundaries, and may convey more clues for us to do recognition.

Fig. 3 shows the idea of DPM in terms of SIFT distributions. From the visible face, we extract the 128-dimensional dense SIFT descriptor from each patch. The descriptors of all patches are then averaged, and the distribution of the 128 values is shown in the middle of Fig. 3(a). This distribution is significantly different from that extracted from the corresponding thermal face, i.e., Fig. 3(c). The DPM transforms the original SIFT distribution into Fig. 3(b), which is much similar to Fig. 3(c) visually. This example clearly illustrates functionality of the DPM.

V. CONCLUSION

In this paper, we present a parametric study to comprehensively investigate performance variations of the deep perceptual model on thermal face recognition, based on various parameter settings and a well-aligned thermal-visible face database. The issues we investigate include how different image descriptors influence performance, how different patch sizes influence performance, how robust the DPM is with different noise levels, and if individuals wearing glasses influence performance. This comprehensive study provides guidelines for future works that adopt DPM as the fundamental method. In the future, more databases will be evaluated, and we will propose variants of DPM to improve recognition performance.

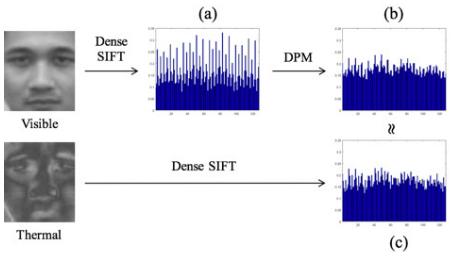

Fig. 3. Illustration of the transformation of the distribution of dense SIFTs by the DPM.

Acknowledgement. This work was partially supported by the Ministry of Science and Technology under the grant 107-2221-E-194-038-MY2 and MOST 107-2218-E-002-054, and the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

REFERENCES

[1] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.

[2] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *Proceedings of International Conference on Biometrics*, 2016.

[3] S. Saxena and J. Verbeek, "Heterogeneous face recognition with cnns," in *Proceedings of ECCV Workshops*, 2016, pp. 483–491.

[4] J. Lezama, Q. Qiu, and G. Sapiro, *Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-Spectral Hallucination and Low-Rank Embedding*, November 2016, https://arxiv.org/abs/1611.06638.

[5] H. Shi, X. Wang, D. Yi, Z. Lei, X. Zhu, and S. Z. Li, "Cross-modality face recognition via heterogeneous joint bayesian," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 81–85, 2017.

[6] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for thermal to visible face recognition," in *Proceedings of British Machine Vision Conference*, 2015.

[7] B. Kresnaraman, D. Deguchi, T. Takahashi, Y. Mekada, I. Ide, and H. Murase, "Reconstructing face image from the thermal infraredspectrum to the visible spectrum," *Sensors*, vol. 16, no. 4, 2016.

[8] A. M. Andres, S. Padovani, M. Tepper, M. Mejail, and J. Jacobo, "Randomized face recognition on partially occluded images," in *Proceedings of Iberoamerican Congress: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 2012.

[9] M. Wang, C. Luo, R. Hong, J. Tang, and J. Feng, "Beyond object proposals: Random crop pooling for multi-label image recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 12, 2016.

[10] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.

[11] X. Chen, P. J. Flynn, and K. W. Bowyer, "Ir and visible light face recognition," *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 332–358, 2005.