# An Examination of Deep-Learning Based Landmark Detection Methods on Thermal Face Imagery

Domenick Poster
West Virginia University
395 Evansdale Dr, Morgantown, WV 26506
dposter@mix.wvu.edu

Shuowen Hu
U.S. Army CCDC Army Research Laboratoryy
2800 Powder Mill Rd, Adelphi, MD 20783
shuowen.hu.civ@mail.mil

Nasser Nasrabadi
West Virginia University
395 Evansdale Dr, Morgantown, WV 26506
nasser.nasrabadi@mail.wvu.edu

Benjamin Riggan
U.S. Army CCDC Army Research Laboratoryy
2800 Powder Mill Rd, Adelphi, MD 20783
benjamin.s.riggan.civ@mail.mil

## Abstract

*Thermal-to-visible face recognition is an emerging technology for low-light and nighttime human identification, for which detection of fiducial landmarks is a critical step required for face alignment prior to recognition. However, thermal images with their low contrast, low resolution, and lack of textural information have proven a challenging obstacle for the detection of the fiducial landmarks used for image alignment. This paper analyzes the ability of modern landmark detection algorithms to cope with the adversarial conditions present in the thermal domain by exploring the strengths and weaknesses of three deep-learning based landmark detection architectures originally developed for visible images: the Deep Alignment Network (DAN), Multi-task Convolutional Neural Network (MTCNN), and a Multi-class Patch-based fully-convolutional neural network (PBC). Our experiments yield a normalized mean squared error of 0.04 at an offset distance of 2.5 meters using the DAN architecture, indicating an ability for cascaded shape regression neural networks to adapt to thermal images. However, we find that even small alignment errors disproportionately reduce correct recognition rates. With images aligned using the best performing model, an 8.2% drop in EER is observed as compared with ground truth alignments, leaving further room for improvement in this area.*

## 1. Introduction

For thermal-to-visible face recognition, faces are aligned to canonical coordinates using a set of fiducial landmarks, which often requires automatic face and landmark detection algorithms. However, there is relatively little research on facial landmark detection in the thermal domain. Moreover, there are indications that face recognition with thermal imagery is more sensitive to proper face alignment compared to visible spectrum face recognition [2]. The alignment process is illustrated in Figure 1.

As stated in [4], "...most of the work to date supports the conclusion that salient facial feature localization in thermal images is significantly more challenging." Thermal imagery inherently has less spatial resolution than visible imagery due to the longer wavelengths of MWIR and LWIR. The facial region in a thermal image exhibits low contrast and lacks the textural information present in its visible counterpart. The modality gap between visible and infrared images is showcased in [6]. Therefore, the plethora of fiducial landmark detection algorithms developed for visible face recognition systems may be challenged in the thermal spectrum.

A variety of deep-learning based approaches have shown success for face alignment on visible images. However, previous thermal alignment research [1][8][14][13][17] has not yet explored deep-learning based approaches. Even with relatively limited amount of training data in the thermal spectrum, the state of the art techniques developed for visible face alignment may still be applicable in the thermal domain through retraining and/or modifications to the network architecture.

In this paper, we examine the importance of face alignment for thermal-to-visible face verification and assess the effectiveness of different landmark detection strategies. We explore the possibility of applying modern deep-learning approaches to the thermal domain. In particular, we investigate a multi-class patch-based fully-convolutional neural network classifier (PBC) and two state of the art landmark

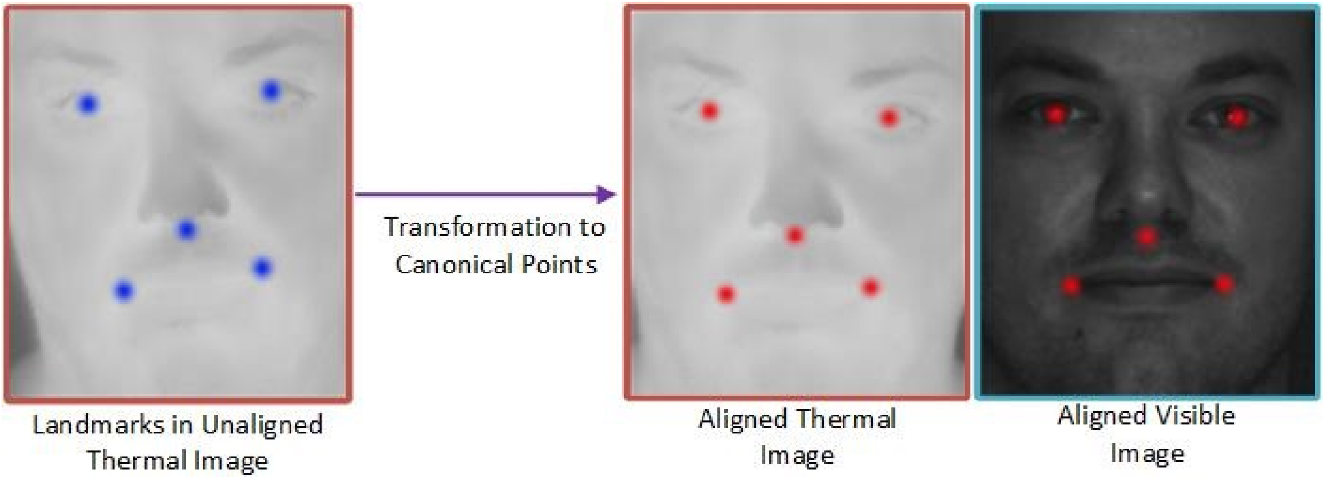
Transformation to Canonical Points
Landmarks in Unaligned Thermal Image
Aligned Thermal Image
Aligned Visible Image

Figure 1. Alignment of a thermal image to a set of canonical points such that the landmarks in both the visible and thermal images appear in the same image locations.

detection algorithms developed for visible images: the Deep Alignment Network (DAN) [9] and the Multi-Task Convolutional Neural Network (MTCNN) [19].

The paper is structured as follows. Chapter 2 summarizes prior research in the area of thermal face alignment. Chapter 3 describes in greater detail the key features of the algorithms examined in this paper. Chapter 4 presents and discusses the experimental results. Closing remarks are made in Chapter 5.

## 2. Background

Early methods of thermal face alignment focused on developing methods specialized to detect a particular facial feature. Bourlai *et al.* [1] relied on a combination of photometric normalization techniques in conjunction with template-based matching to detect eye locations. Also focusing on eye detection, Wang *et al.* [17] extract Haar-like features from assumed eye regions for classification with an SVM. Utilizing video frames to detect nostril locations, Tzeng *et al.* [14] exploit the variance in temperature around the nostrils during respiration.

Recent research targets the facial region as a whole. Kopaczka *et al.* [8] learn an Active Appearance Model from HOG and SIFT features to track landmark locations in thermal videos. Bypassing facial landmarks altogether, Sun and Zheng [13] perform iterative point-to-point matching with Canny edge maps of visible and thermal image pairs.

In 2003, Chen *et al.* [2] studied the sensitivity of combined thermal and visible face recognition systems to images misaligned via small perturbations in the eye landmark locations. They recorded a 5-19% drop in correct match percentages compared to using correctly aligned images. It should be noted the systems used both the visible and thermal image of a probe subject, as opposed to using only a thermal probe. The algorithms consisted of a PCA-based matcher as well as a commercial off-the-shelf system.

## 3. Landmark Detection in Thermal Imagery

### 3.1. Multi-Task Convolutional Neural Network

MTCNN [19] is a joint face detection and landmark localization algorithm used in the preprocessing pipeline of several state of the art face recognition models [3][10][16]. The architecture is composed of a three stage neural network. Each stage of the network is trained to simultaneously classify face regions and directly regress a set of landmark location values for each region.

The *i*th stage is defined as

$$\hat{y}_i^{face}, \hat{y}_i^{box}, \hat{y}_i^{l} = f_i(P;\theta), \tag{1}$$

where $P$ is an image patch, $\hat{y}_i^{face} \in \mathbb{R}$ is the probability that $P$ is a face, $\hat{y}_i^{box} \in \mathbb{R}^4$ are adjustments to the position and size of the bounding box describing the face patch, $\hat{y}_i^{l} \in \mathbb{R}^{10}$ is the set of coordinates for the eyes, nose, and mouth corner landmarks. Throughout the paper, network parameters are denoted as $\theta$. Each stage is trained independently, with input batches alternating the minimization of the losses associated with $\hat{y}^{face}$, $\hat{y}^{box}$, and $\hat{y}^{l}$. Input training data is composed of image patches of size 12x12, 24x24, and 48x48 for stages $i = 1...3$ respectively. Positive (face) patches and randomly cropped negative (non-face) patches are used in the training of the face classification task.

After training, the first stage acts as a fully-convolutional network which produces a set of feature maps, where each spatial location in the output feature maps is a vector containing $\hat{y}^{face}$, $\hat{y}^{box}$, and $\hat{y}^{l}$ associated with a specific receptive field. The location of face regions in the original input image can be extrapolated from the feature maps based on the receptive field of the network. Detected face regions are cropped and propagated to the next stage. Stages 2 and 3, operating on proposed patches of incrementally larger resolutions, yield $1 \times 1 \times 15$ dimensional feature maps corresponding to the 10-dimension landmark values, the 4-dimensional bounding box values, and the 1-dimensional face class probability. The stages of the MTCNN resemble a cascaded, coarse-to-fine detection strategy.

| **layer** | **size** | **kernel** | **stride** | **padding** |
|---|---|---|---|---|
| conv | 10 | 3x3 | 1 | valid |
| max pool |  | 2x2 | 2 | same |
| conv | 16 | 3x3 | 1 | valid |
| conv | 32 | 3x3 | 1 | valid |
| face | 1 | 1x1 | 1 | valid |
| box | 4 | 1x1 | 1 | valid |
| landmarks | 10 | 1x1 | 1 | valid |

Table 1. Stage 1 of MTCNN.

| **layer** | **size** | **kernel** | **stride** | **padding** |
|---|---|---|---|---|
| conv | 28 | 3x3 | 1 | valid |
| max pool |  | 3x3 | 2 | same |
| conv | 48 | 3x3 | 1 | valid |
| max pool |  | 3x3 | 2 | valid |
| conv | 64 | 2x2 | 1 | valid |
| max pool |  | 2x2 | 2 | valid |
| fc | 128 | NA | NA | NA |
| face | 1 | 1x1 | 1 | valid |
| box | 4 | 1x1 | 1 | valid |
| landmarks | 10 | 1x1 | 1 | valid |

Table 2. Stage 2 of MTCNN.

| **layer** | **size** | **kernel** | **stride** | **padding** |
|---|---|---|---|---|
| conv | 32 | 3x3 | 1 | valid |
| max pool |  | 3x3 | 2 | same |
| conv | 64 | 3x3 | 1 | valid |
| max pool |  | 3x3 | 2 | valid |
| conv | 64 | 3x3 | 1 | valid |
| max pool |  | 2x2 | 2 | same |
| conv | 128 | 2x2 | 1 | valid |
| fc | 256 | NA | NA | NA |
| face | 1 | 1x1 | 1 | valid |
| box | 4 | 1x1 | 1 | valid |
| landmarks | 10 | 1x1 | 1 | valid |

Table 3. Stage 3 of MTCNN.

Tables 1, 2, and 3 present the network architecture of MTCNN's three stages. The layers entitled "face", "box", and "landmarks" represent the $\hat{y}^{face}$, $\hat{y}^{box}$, and $\hat{y}^{l}$ outputs of the network and are each connected to the the last convolutional or fully-connected layer of the network. The three stages of MTCNN contain a total of 494,924 paramaters, however the majority of parameters exist in the final stage (387,648 parameters).

Because the MTCNN jointly regresses all five landmarks from the entire face region, it learns a spatial arrangement of facial features, leading to anatomically reasonable guesses when there is a lack of information to track individual landmarks.

### 3.2. Multi-Class Patch-Based Classifier

The Multi-Class Patch-Based Classifier (PBC) detects facial features similar to how the MTCNN detects faces. Where MTCNN's face detector is a binary classifier, the PBC classifies regions of an image as belonging to one of six classes, five of which correspond to landmark locations (left eye, right eye, base of the nose, left mouth corner, and right mouth corner) while the sixth represents a non-landmark region.

The Multi-Class PBC is a fully-convolutional neural network trained on image patches extracted from landmark and non-landmark facial regions, similar to MTCNN and other cascaded classifiers such as Viola-Jones [15]. The architecture is constructed such that a 60x60 input image patch becomes spatially reduced through the network such that the output is a single vector of class probabilities. The structure of the network, detailed in Table 4, is based off the final stage of MTCNN. The network has 1,016,390 parameters.

| **layer** | **size** | **kernel** | **stride** | **padding** |
|---|---|---|---|---|
| conv | 32 | 3x3 | 1 | valid |
| max pool |  | 3x3 | 2 | same |
| conv | 64 | 3x3 | 1 | valid |
| max pool |  | 3x3 | 2 | valid |
| conv | 64 | 3x3 | 1 | same |
| max pool |  | 2x2 | 2 | valid |
| conv | 128 | 2x2 | 1 | same |
| conv | 256 | 3x3 | 1 | same |
| conv | 256 | 3x3 | 1 | same |
| conv | 10 | 1x1 | 1 | same |

Table 4. Multi-Class Patch-Based Classifier (PBC) network.

After training, the network is fed a cropped face image $I \in \mathbb{R}^{h \times w}$ and produces a three-dimensional feature map $M \in \mathbb{R}^{j \times k \times c}, \{(j|k|c) : j < h, k < w, c = 6\}$ of $c$ of unscaled class logits. The PBC network function is defined as

$$M = f(I; \theta). \tag{2}$$

The indices in $M$ with the highest classification score for class $l$ are given by

$$\hat{x}_l, \hat{y}_l = \underset{j,k}{\operatorname{argmax}}(M_{j,k,l}). \tag{3}$$

Given a function $g(p, q)$ mapping from spatial location $(p, q)$ in $M$ to region $R \in \mathbb{R}^{n \times m}$, $\{(n|m) : n < h, m < w\}$ in $I$. The region of $I$ containing landmark $l$ is given by

$$R_l = g(\underset{j,k}{\operatorname{argmax}}(M_{j,k,l})). \tag{4}$$

The $(x, y)$ coordinate of the landmark in the original image $I$ is assumed to be the center point of $R_l$.

In contrast to the MTCNN, PBC classifies each region of the image independently, paying no regard to the global appearance. As the classifier focuses entirely on local regions, its parameters become specialized for the detection of specific features. However, false positives can lead to large errors since the model is not constrained by a global face shape prior.

### 3.3. Deep Alignment Network

DAN [9] is a state of the art landmark detection algorithm for visible images. It is composed of two cascaded VGG-like networks. A 112x112 input image $I$ is aligned to an initial estimate of the landmarks $l$, usually obtained from a mean shape calculated from the training data. Whereas MTCNN regresses landmark location values, DAN regresses a set of offsets used to update the initial landmark estimates. Similar to MTCNN, DAN learns a statistical representation of a face by regressing landmark values from the global image.

Our experiments utilize the two stage version of the model. The first stage is defined simply as

$$\Delta l_1 = S_1(I; \theta). \tag{5}$$

The input to the second stage is a three channel image composed of the original image, a heatmap image $H$ highlighting estimated landmark locations, and a feature embedding vector $E$ obtained from a fully-connected layer in the prior stage. Between stages, a similarity transform $T$ is used to re-normalize the image to the canonical shape $l$. The largest of the three models, the two-stage DAN contains 23,022,592 total parameters.

$$\Delta l_2 = S_2(T(I), H, E; \theta). \tag{6}$$

## 4. Experimental Results

### 4.1. Dataset

This study uses Volumes 1 and 2 of the ARL Polarimetric Thermal Face Dataset released in [7] and extended in [18]. It is a collection of thermal and visible image pairs. Volume 1 contains 60 subjects while Volume 2 contains 51 subjects.

Data from Volume 1 is captured at three different distances: Range 1 (2.5m), Range 2 (5m), and Range 3 (7.5m). Each of the 60 subjects has 16 image samples per range, for a total of 48 samples per subject. The average interocular distances for the thermal images are 87 pixels, 44 pixels, and 31 pixels at Ranges 1, 2, and 3 respectively. The interocular distance of the visible images at Range 2 matches the interocular distance for Range 1 thermal images. Volume 2 data is captured at Range 1 only (2.5m), with 31 samples per subject.

The dataset is divided into Protocols 1 and 2 as described in [18]. For both Protocols, five random folds are generated wherein subjects are randomly assigned to train and test sets. The 60 subjects in Protocol 1 are evenly split between training and testing. Each fold of Protocol 2 is created by randomly selecting 85 subjects for training and 26 subjects for testing.

Images are horizontally flipped to augment the training data. As a result, Protocol 1 contains 23,040 training images and 11,520 testing images per fold. Protocol 2 contains on average 72,845 training images and 11,098 testing images per fold.

### 4.2. Training

The DAN and MTCNN models have been trained in the same fashion as described in the original papers [9][19].

Because MTCNN performs face detection and landmark localization jointly, it is possible for it to fail to propose the correct face region, thereby failing to provide any landmarks. In contrast, DAN and PBC assume a cropped face is given. In order to facilitate a fair comparison, the first two stages of MTCNN are bypassed at test time, which serve only to propose and refine the detected face region. Instead, the correct face region is passed directly to the third stage to obtain the regressed landmark values.

The PBC is trained for four epochs on randomly cropped landmark and non-landmark locations. Random patches are considered to be landmark regions if they have an Intersection over Union greater than 0.8 with the ground truth landmark region. A learning rate of 0.001 is used with Adam optimizer on batch sizes of 128.

### 4.3. Landmark Detection Performance

The algorithms are evaluated based on the point-to-point normalized mean squared error (NMSE) metrics, calculated

| Method | R1 | R2 | R3 | Avg |
|---|---|---|---|---|
| MTCNN | $0.201 \pm 0.02$ | $0.205 \pm 0.02$ | $0.187 \pm 0.03$ | $0.198 \pm 0.03$ |
| PBC | $0.164 \pm 0.02$ | $0.330 \pm 0.06$ | $0.716 \pm 0.05$ | $0.403 \pm 0.03$ |
| DAN | $0.050 \pm 0.02$ | $0.056 \pm 0.02$ | $0.061 \pm 0.02$ | $0.056 \pm 0.02$ |

Table 5. Protocol 1 NMSE and standard deviations at Ranges 1, 2, and 3.

| Method | R1 | R2 | R3 | Avg |
|---|---|---|---|---|
| MTCNN | $0.112 \pm 0.03$ | $0.109 \pm 0.03$ | $0.114 \pm 0.03$ | $0.097 \pm 0.03$ |
| PBC | $0.073 \pm 0.01$ | $0.117 \pm 0.02$ | $0.236 \pm 0.06$ | $0.211 \pm 0.04$ |
| DAN | $0.044 \pm 0.02$ | $0.045 \pm 0.02$ | $0.047 \pm 0.02$ | $0.046 \pm 0.02$ |

Table 6. Protocol 2 NMSE and standard deviations at Ranges 1, 2, and 3.

as

$$e = \frac{\|l_i - l^*\|_2}{d} \tag{7}$$

where $d$ is the interocular distance, $l_i$ is the $i$th predicted landmark, and $l^*$ is the ground truth landmark.

Tables 5 and 6 list the global NMSE and standard deviation averaged over all landmarks across all five folds of each protocol. DAN achieves the lowest error rates at all ranges and in both protocols. Figures 2 and 3 plot the Cumulative Error Distribution (CED) curves, representing the proportion of images whose average NMSE falls below a given error threshold.

Protocol 2 contains more than triple the amount of training images than Protocol 1. The limited training data of Protocol 1 results in substantially lower performance across all algorithms. DAN and MTCNN perform consistently across all ranges due to their ability to learn a holistic representation of a face. Conversely, the lower resolutions at increased ranges causes PBC to suffer drastically.

Further insight into the algorithms' behaviors are gained from the qualitative examples in Figure 4. The example image exhibits how PBC misclassifies the left eye corner as the left mouth corner. The plateu regions of the PBC's CED curve highlight the negative impact Ranges 1 and 2 have on its performance. While PBC failed to outperform DAN, a form of local region refinement may improve global appearance-based methods in the case of high-resolution face images

The right-shifted CED curve for MTCNN is an indication of the consistent amount of error being introduced to each landmark. This is reflected in the central image of Figure 4. By regressing exact landmark locations, the range and scale of output values is larger for MTCNN than DAN. Outputs in this range may be harder to control than DAN's smaller, iterative shape updates. The fact that MTCNN occasionally sees slightly lower NMSE at some higher ranges points to some predictions being coincidentally accurate.

Additional qualitative results in Figure 5 showing the performance for each model on each of three subjects re-

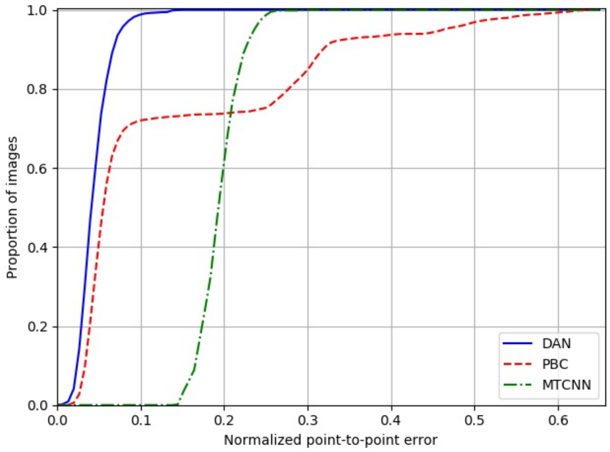

Figure 2. Protocol 1 Cumulative Error Distribution curves of the NMSE for DAN, PBC, and MTCNN.

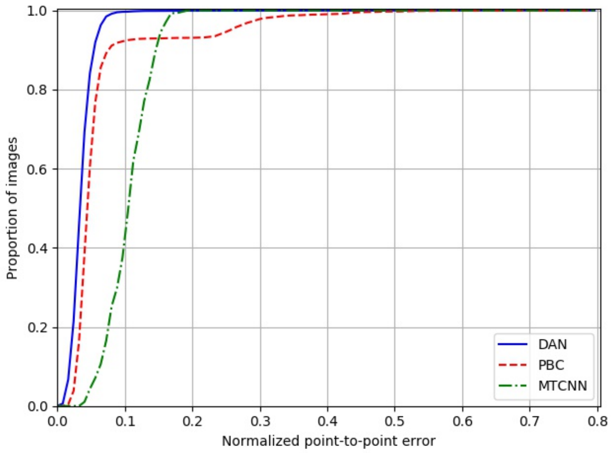

Figure 3. Protocol 2 Cumulative Error Distribution curves of the NMSE for DAN, PBC, and MTCNN.

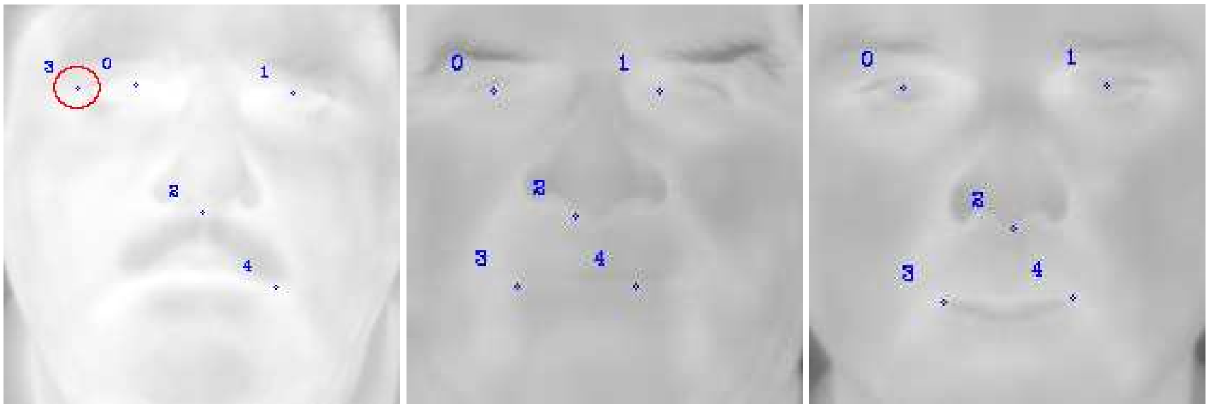

Figure 4. Left: PBC landmark predictions containing a corner of the left eye being mistaken for the left mouth corner. Center: MTCNN landmarks with correct spatial arrangement but wrong locations. Right: High quality DAN landmarks.

inforce the previous qualitative observations and corroborate the quantitative results. Despite the predictions of the PBC model being on par with or better than predictions from the DAN model for the subjects in the first two columns, the PBC's failure to accurately localize the right mouth corner for the third subject heavily skews its average NMSE, once again insinuating some form of local refinement may be beneficial in tandem with global information.

### 4.4. Impact on Face Verification

We follow the same process as [7][11] for conducting face verification trials, however we align to five points instead of two. DoG filtering is applied to the aligned thermal and visible imagery. As in [7][11], a Deep Perceptual Mapping (DPM) [12] from visible to thermal modalities is learned. Finally, matching is performed with one-versus-all classifiers using a partial least squares (PLS) regression model [5].

The following results are verification rates for Range 1 thermal image probes. Figures 6 and 7 illustrates the drop in performance when using the predicted landmarks for verification versus the ground truth. However, DAN landmarks nearly match performance with the ground truth on Protocol 2. This demonstrates the ability of DAN to adapt effectively to thermal imagery.

Taken as a whole, the results concur with the findings of [2] that face verification in the thermal spectrum is more sensitive to image alignment.

## 5. Conclusion

Our results illustrate the sensitivity of face verification algorithms to misaligned thermal images. We have shown that thermal images aligned with modern landmark detection algorithms often fail to achieve thermal-to-visible face verification results on par with manually aligned imagery.

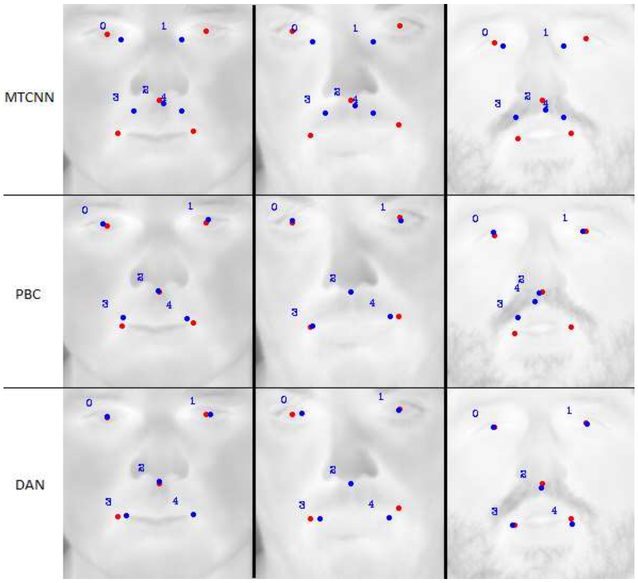

Figure 5. Qualitative results for each model on Protocol 1 at Range 1 (2.5m). The enumerated detected landmarks are shown in blue, ground truth landmarks in red.

Nevertheless, we demonstrate the cascaded shape regression method exhibited by the DAN architecture shows promise. Learning a global face appearance is key to avoiding critical localization errors, especially at offset distances from the camera greater than 2.5 meters. However, quantitative findings hint at the potential benefits of integrating local and global detection strategies when high resolution, high inter-ocular distance thermal images are available.

The benefits of joint face detection and landmark localization exploited by MTCNN for visible images does not appear to translate to the thermal domain, where there may be more benefit for algorithms to specialize in accomplishing a single task given the unique qualities of thermal im-

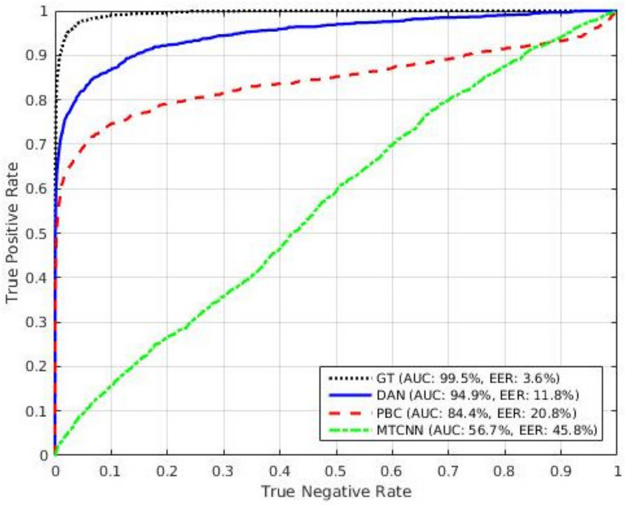

Figure 6. Protocol 1 ROC curves showing thermal-to-visible verification performance when using ground truth (GT), DAN, PBC, and MTCNN landmarks.

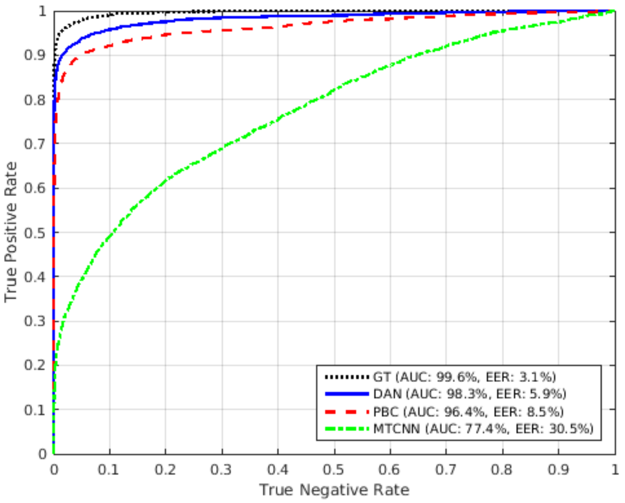

Figure 7. Protocol 2 ROC curves showing thermal-to-visible verification performance when using ground truth (GT), DAN, PBC, and MTCNN landmarks.

agery. Another important characteristic is that while both MTCNN and DAN represent regression-based strategies, DAN's approach of iteratively regressing landmark updates, instead of the actual landmark coordinates, is likely an easier objective to learn.

## Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-18-2-0145. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## References

[1] Thirimachos Bourlai and Zain Jafri. Eye detection in the middle-wave infrared spectrum: towards recognition in the dark. In *Information Forensics and Security (WIFS), 2011 IEEE International Workshop on*, pages 1–6. IEEE, 2011. 1, 2

[2] Xin Chen, Patrick J Flynn, and Kevin W Bowyer. Visible-light and infrared face recognition. In *Workshop on Multimodal User Authentication*, page 48. Citeseer, 2003. 1, 2, 6

[3] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 2

[4] Reza Shoja Ghiass, Ognjen Arandjelović, Abdelhakim Bendada, and Xavier Maldague. Infrared face recognition: A comprehensive review of methodologies and databases. *Pattern Recognition*, 47(9):2807–2824, 2014. 1

[5] Shuowen Hu, Jonghyun Choi, Alex L Chan, and William Robson Schwartz. Thermal-to-visible face recognition using partial least squares. *JOSA A*, 32(3):431–442, 2015. 6

[6] Shuowen Hu, Nathaniel Short, Benjamin S Riggan, Matthew Chasse, and M Saquib Sarfraz. Heterogeneous face recognition: recent advances in infrared-to-visible matching. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 883–890. IEEE, 2017. 1

[7] Shuowen Hu, Nathaniel J. Short, Benjamin S. Riggan, Christopher Gordon, Kristan P. Gurton, Matthew Thielke, Prudhvi Gurram, and Alex L. Chan. A polarimetric thermal database for face recognition research. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 4, 6

[8] Marcin Kopaczka, Kemal Acar, and Dorit Merhof. Robust facial landmark detection and face tracking in thermal infrared images using active appearance models. In *VISIGRAPP (4: VISAPP)*, pages 150–158, 2016. 1, 2

[9] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, volume 3, page 6, 2017. 2, 4

[10] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 1, 2017. 2

[11] Benjamin S Riggan, Nathaniel J Short, and Shuowen Hu. Optimal feature learning and discriminative framework for

polarimetric thermal to visible face recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–7. IEEE, 2016. 6

[12] M Saquib Sarfraz and Rainer Stiefelhagen. Deep perceptual mapping for cross-modal face recognition. *International Journal of Computer Vision*, 122(3):426–438, 2017. 6

[13] Lin Sun and Zengwei Zheng. Thermal-to-visible face alignment on edge map. *IEEE Access*, 5:11215–11227, 2017. 1, 2

[14] Huan-Wen Tzeng, He-Chin Lee, and Mei-Yung Chen. The design of isotherm face recognition technique based on nostril localization. In *System Science and Engineering (ICSSE), 2011 International Conference on*, pages 82–86. IEEE, 2011. 1, 2

[15] Paul Viola, Michael Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1:511–518, 2001. 3

[16] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *arXiv preprint arXiv:1801.09414*, 2018. 2

[17] Shangfei Wang, Zhilei Liu, Peijia Shen, and Qiang Ji. Eye localization from thermal infrared images. *Pattern Recognition*, 46(10):2613–2621, 2013. 1, 2

[18] He Zhang, Benjamin S Riggan, Shuowen Hu, Nathaniel J Short, and Vishal M Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *arXiv preprint arXiv:1812.05155*, 2018. 4

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 2, 4