TFLD: Thermal Face and Landmark Detection for Unconstrained Cross-spectral Face Recognition

David Anghelone, Sarah Lannes, Valeriya Strizhkova, Philippe Faure, Cunjian Chen, Antitza Dantcheva

TFLD: Thermal Face and Landmark Detection for Unconstrained Cross-spectral Face Recognition

David Anghelone[1,2,3], Sarah Lannes[2], Valeriya Strizhkova[1,3], Philippe Faure[2], Cunjian Chen[4,5] and Antitza Dantcheva[1,3]

[1]Inria [2]Thales [3]Univsersité Côte d'Azur [4]Monash University [5]Monash Suzhou Research Institute

Abstract

Automated thermal-to-visible face recognition has received increased attention due to benefits related to low-light applications. Towards improvement of related matching accuracy, we hereby present TFLD, a detector of face and landmarks operating in the thermal spectrum. Our proposed TFLD is based on the architecture of YOLOv5, integrating sequential modules for face and landmark detection. We introduce a thermal face restoration scheme, in order to enhance thermal image quality and hence detection accuracy. We address data scarcity by transferring landmarks in paired visible and thermal images. Our experimental results showcase that our proposed detector accurately detects faces, as well as landmarks in a wide range of adversarial conditions. Further, TFLD achieves promising results on three benchmark multi-spectral face and landmark datasets, namely ARL-VTF, SF-TL54 and RWTH-Aachen; thereby improving the matching accuracy in cross-spectral face recognition by providing robust face alignment based on estimated facial landmarks.

1. Introduction

Face recognition (FR) generally includes *face detection* and landmark-based *alignment* as initial processing steps [6, 2, 13, 17]. While *landmark detection* has become reasonably reliable in the context of the visible spectrum, it remains challenging in the context of thermal images due to associated inherent *low-contrast and low-resolution*, as well due to *poor texture information* [16]. We note that existing face and landmark detection algorithms, that were trained with visible face images, fail to generalize onto thermal images [7, 14] due to the cross-spectral *modality gap*. At the same time, *lack of* available *annotated thermal datasets* is the primary cause for the scarcity of work focused on detection of thermal facial landmarks [8, 11, 16].

Motivated by the above, we present a novel *thermal face* and *landmark detector (TFLD)*, streamlined to be *robust to* adversarial conditions such as *pose, expression, occlusion, poor image quality and long-range distance*. Specifically jointly, TFLD and a proposed data augmentation strategy, are able to (i) detect face and landmarks in the thermal spectrum in challenging unconstrained conditions. Related to that, TFLD (ii) establishes a benchmark for face and landmark detection in the thermal spectrum. We present related results on the ARL-VTF [16] dataset and further showcase that TLFD is instrumental for (iii) cross-spectral face recognition (CFR), which aims to compare visible face images against face images acquired beyond the visible, as well as to (iv) assist thermal monitoring systems (see Figure 1). In addition, TFLD (v) enhances currently limited annotation of existing thermal face datasets, e.g., 5 facial landmarks in the ARL-VTF dataset by detecting 68 facial landmarks. However, when datasets contain rich landmark annotations, such as SF-TL54 [10] or RWTH-Aachen [8] datasets, TFLD training can be directly applied. To be specific to the ARL-VTF database, TFLD extracts facial landmarks in visible face images and transfers these to the synchronized and aligned thermal counterpart images, in order to serve as ground truth annotations.

Given that face detection and facial landmarks constitute sub-tasks of traditional object detection, we hereby adopt YOLOv5[1], which has excelled in object detection. Therefore, TFLD detects facial landmarks by considering them as the center of a textured area instead of points. In contrast to visible spectrum images, thermal images contain less high-frequency information and associated degraded quality renders semantic definitions for certain landmarks challenging. To the best of our knowledge, this is the first work based on YOLOv5 for large-scale thermal-based facial landmark detection. Our method brings more benefits compared to prior work, in particular offering, on the one hand, the ability to detect a large amount of thermal facial landmarks in unconstrained environment. And on the other hand, providing facial key points for the purpose of face alignment, which also demonstrates a positive impact on the face recognition

[1]https://github.com/ultralytics/yolov5

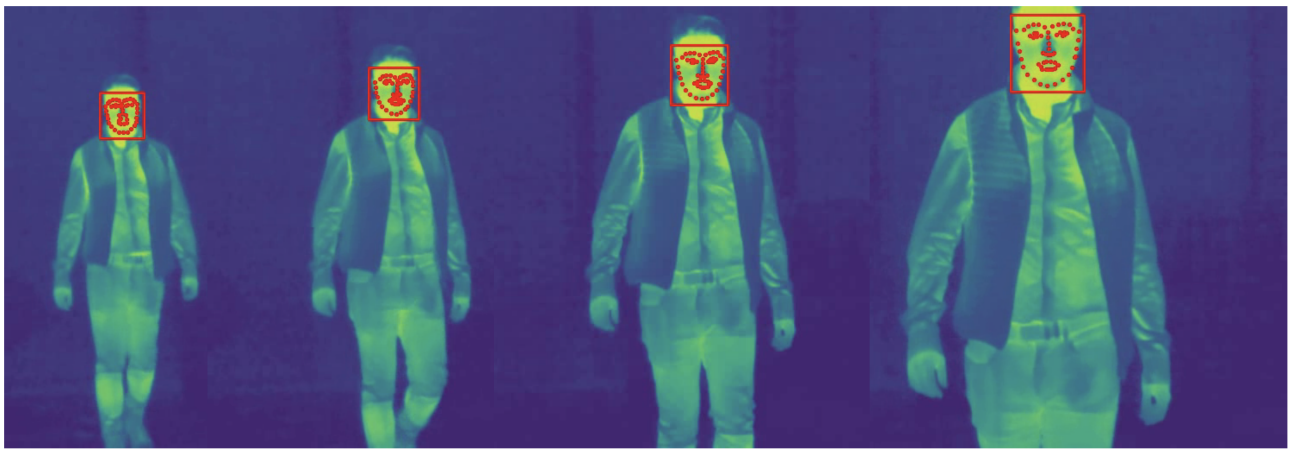

Figure 1. **Monitoring system with thermal sensor.** TFLD method applied on video sequence captured in the wild. A person, approximately 14m away, walks towards the camera while TFLD is tracking face and landmarks.

scores, therefore rendering TFLD an accurate automatic annotation tool for cross-spectral face recognition systems.

We design a model, where a thermal face restoration (TFR) pre-processing filter is succeeded by two YOLOv5 models, denoted as $M1$ and $M2$. TFR is beneficial in highlighting visual details and contours, in improving contrast and sharpness of the face, ultimately allowing for better detection accuracy. While $M1$ detects the full face in the thermal spectrum, $M2$ subsequently detects a set of facial landmarks in the localized face. We evaluate the accuracy of TFLD by assessing landmark accuracy, as well as by determining the impact of the proposed face alignment on CFR.

The main contributions of this work include the following.

- We propose a novel framework incorporating two successive YOLOv5-object detectors for face and landmark detection in the thermal spectrum, placing emphasis on robustness in unconstrained environments. TFLD predicts landmarks as regions of interest instead of specific marks, where textured areas are a principal concern rather than semantic points. We incorporate a thermal face restoration module as a pre-processing filter, allowing for a significant detection improvement.

- We (a) achieve at least state-of-the-art performance on three benchmark thermal face datasets with respect to *landmark localisation* and (b) improve automatic face recognition *matching scores* when using TFLD for pre-processing.

The rest of the paper is organized as follows. Section 2 revisits recent work on landmark detection involving thermal imaging. Section 3 introduces the framework of the proposed TFLD. Section 4 presents experimental results pertaining to face and landmark detection, as well as the impact of TFLD based face alignment on CFR. Section 5 concludes the work.

2. Related Work

A number of deep learning based approaches have been proposed to address the task of face landmark detection in the thermal spectrum. Poster et al. [14] examined three different approaches, namely deep alignment network (DAN), multi-task convolutional neural network (MTCNN) and multi-class patch-based fully convolutional neural network (PBC). Similarly, Chu and Liu [4] presented a neural network approach for joint facial landmark detection and emotion recognition. Kuzdeuov et al. [10] compared the classical machine learning model Dlib based on a set of regression trees with a deep learning model based on the U-net architecture. All these approaches were originally designed for visible face landmark detection and were then retrained for thermal images. However, the large inter-spectral gap left room for improvement with respect to detection results. In addition, face and landmark detection in the thermal spectrum was addressed by methods based on generative adversarial networks (GANs), aiming at translating facial images to the visible spectrum, then extracting facial key points and transferring the key points to the image of the original spectrum. Mallat et al. [11] proposed converting existing visible face databases to the thermal spectrum. Active appearance models (AAMs) and DAN are then trained using the synthetic thermal, along with the shared landmark annotations. Nevertheless, Nagumo et al. [12] noted that computation using AAM method is costly, thus rendering AAM inapplicable in real-world scenarios. Poster et al. [15] leveraged the use of visible data by proposing visible-to-thermal parameter transfer learning with a coupled neural network. However, Anghelone et al. [1] reported that the facial identity was often not preserved dur-

ing the spectral transformation, where salient regions of the synthesized face deviated. As a result, the methodology frequently failed in facial landmark localization.

3. Proposed Approach

Our proposed TFLD is based on YOLOv5, comprised of two sequential modules, namely a face detection module and a landmark detection module. Prior to these, a thermal face restoration pre-processing filter is applied to enhance the thermal image quality before the detection. The overall architecture of TFLD is illustrated in Figure 4.

3.1. Formalization

Let $\mathcal{T}$ be the thermal domain. A digital thermal image can be defined as

$$\begin{array}{rcccc} I^{thm}_{w\times h} & : & \{0,..,w-1\}\times\{0,..,h-1\} & \rightarrow & \{0,..,255\} \\ & & (x,y) & \mapsto & I^{thm}_{w\times h}(x,y), \end{array} \tag{1}$$

where $w \in \mathbb{N}^*$ and $h \in \mathbb{N}^*$ denote the width and height, respectively. To simplify the notation, we denote $I^{thm}_{w\times h,n}$ for a thermal image belonging to $\{I^{thm}_{w\times h,n}\}^N_{n=1}$, which represents a set of N thermal images. Equation (1) defined as *thermogram* is the function, which quantifies the light intensity ranging from 0 to 255, emerging through the heat sensitive energy acquired from any point of the thermal sensor. Due to the nature of thermal imagery, a face will emit significant heat and thus will appear as a high intensity object. This physical phenomenon is illustrated in Figure 2 (left).

3.2. Thermal face restoration

Motivated by the fact that thermal sensors provide both, poor image quality and low spatial resolution, we here note that improvement of the former is essential for accurate facial landmark locatalization. We introduce a thermal face restoration (TFR) pre-processing filter that reveals many visual details, enhancing contrast and sharpness [3]. TFR is based on a combination of several difference of Gaussians (DoG) filters. The DoG filter as main operation, noted $\Gamma_{\sigma_i,\sigma_j}$, serves as a spatial band-pass filter and involves the subtraction of one Gaussian blurred G_{σ_i} version of an original image from another G_{σ_j} less blurred version of the original. The Gaussian blurred image B_σ is obtained by convoluting the original thermal image $I^{thm}_{w\times h}$ with a Gaussian kernel G_σ having a standard deviation σ. This can be expressed as

$$B_\sigma = I^{thm}_{w\times h} * G_\sigma, \tag{2}$$

where the Gaussian kernel is from a two-dimensional Gaussian distribution

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{3}$$

Figure 2. **Thermal face restoration** pre-processing filter. Left image shows the raw image $I^{thm}_{w\times h}$ acquired from a thermal sensor, whereas right image shows the enhanced image TFR($I^{thm}_{w\times h}$) performed by the TFR filter.

In particular, the DoG filter is obtained by performing the subtraction of two Gaussian kernels. Here, a kernel must have a standard deviation σ_i lower than the previous σ_j.

$$\Gamma_{\sigma_i,\sigma_j} = B_{\sigma_j} - B_{\sigma_i}. \tag{4}$$

Finally, the TFR image is obtained via a combination of two DoG filters consisting of both, different Gaussian kernel size and standard variation values, and can be formalized as

$$TFR(I^{thm}_{w\times h}) = \Gamma_{\sigma_1,\sigma_2} + \Gamma_{\sigma_3,\sigma_4}. \tag{5}$$

Figure 2 highlights quality enhancement, as contributed by TFR.

3.3. Data augmentation

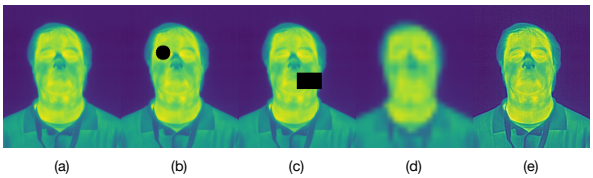

Figure 3. **Data augmentation.** An original image (a) augmented by introducing (b) circular occlusion, (c) rectangular occlusion, (d) low resolution degradation, and (e) thermal face restoration processing.

We augment our dataset with a set of simulations of real world conditions such as occlusions (e.g., random circles and rectangles superimposed on the image), low resolution degradation and long range acquisition variations, see Figure 3. Such simulations allow the model to gain robustness to unconstrained environments, and hence to adapt to in-the-wild scenario.

3.4. Missing ground truth extrapolation with visible-to-thermal landmark transfer

Thermal face datasets often lack labeled face bounding boxes and offer limited corresponding landmarks, e.g., of the left and right eyes, nose and left and right mouth corners only. As a result, very limited scientific work has been

focusing on thermal facial landmark detection. For our purpose, we consider the ARL-VTF [16] dataset, introduced in Section 4.1.1. It is a large scale dataset including synchronized and aligned visible-thermal face images offering only a partial set of thermal landmarks. We extract facial landmarks based on visible images and proceed to transfer these onto the thermal counterpart, in order to employ them as ground truth references. This is illustrated in Figure 5, providing a full facial landmark annotation to the images in the thermal spectrum.

3.5. Baseline model

We design a series of two successive YOLOv5 models based on the medium backbone, denoted as $M1$ and $M2$ respectively, where both are optimized by three objective functions that include (i) the mean square error as *bounding box regression loss*, (ii) the binary cross entropy as *objectness loss* and (iii) the cross entropy as *classification loss*. While $M1$ detects the regions of interest (ROI) that contains faces from the background, $M2$ aims at extracting a set of landmarks pertaining to the prior cropped region. Given that YOLOv5 was originally an object detector, we build $M1$ and $M2$ for face and landmark detection, rather than box detection. In order to do so, we consider (a) a face F as a bounding area based on the semantic definition of the inter-eye distance d_{IED} (center of the left and right eyes). F contains at least the left and right eyes, nose and mouth. In particular, we apply the following standard where the upper left coordinates (up) and the bottom right coordinates (down) of F are defined as

$$(x_{up}, y_{up}) = (x_{\text{ left eye}} - \frac{d_{IED}}{2}, y_{\text{ left eye}} - \frac{d_{IED}}{3}), \quad (6)$$

$$(x_{down}, y_{down}) = (x_{\text{ right eye}} + \frac{d_{IED}}{2}, y_{\text{ right eye}} + \frac{2d_{IED}}{3}). \quad (7)$$

In addition, we consider (b) a landmark l_k as a ROI through a custom box, where the center represents the desired landmark with proportional width and height shape. Consequently, for all $k \in [0, K]$ the associated landmark l_k is expressed by the shape of the custom box with

$$l_k = I^{thm}_{w\times h}(x_k, y_k) = (\frac{x^k_{up} + x^k_{down}}{2}, \frac{y^k_{up} + y^k_{down}}{2}), \quad (8)$$

where K is the total number of facial landmarks, (x^k_{up}, y^k_{up}) and (x^k_{down}, y^k_{down}) denote the upper left and bottom right coordinates corner of the k-th custom box, respectively.

The TFR filter is first applied as a pre-processing step on the n-th thermal image $I^{thm}_{w\times h,n}$. Hence, the network is fed by $TFR(I^{thm}_{w\times h,n})$, where $M1$ detects a set of F_n faces. If no face is detected, $M1$ returns an empty set $\emptyset$. This is formalized as follows.

$$M1(TFR(I^{thm}_{w\times h,n})) = \begin{cases} \emptyset & \text{if } F_n = 0 \\ \bigcup_{f=1}^{F_n} F^f_{w_c\times h_c,n} & \text{otherwise,} \end{cases} \quad (9)$$

where $f \in [1, F_n]$ and $F^f_{w_c\times h_c,n}$ is the f-th cropped face[2] within the n-th image.

Then, given the Equation (9), the second model $M2$ produces the final landmark output. Hence, for $F_n \neq 0$ and all $f \in [1, F_n]$, $M2$ attempts to provide a set $L_{f,n}$ of K landmarks corresponding to the f-th face of the n-th image. This is formalized as follows.

$$M2(F^f_{w_c\times h_c,n}) = L_{f,n}. \quad (10)$$

In particular,

$$L_{f,n} = \bigcup_{k=1}^{K} l^{f,n}_k, \quad (11)$$

where $l^{f,n}_k$ refers to the coordinates $(x^{f,n}_k, y^{f,n}_k)$ of the k-th landmark present in the f-th face of the n-th image. If $l^{f,n}_k$ is undetected, $M2$ returns an empty point $\emptyset$, which then interpolated with other based predicted landmarks.

Finally, for all $f \in [1, F_n]$, $L_{f,n}$ is reported on the original $I^{thm}_{w\times h,n}$ thermal input image, providing therefore the final landmark locations.

4. Experimental Results

4.1. Dataset and Protocol

4.1.1 Thermal face datasets

To validate the effectiveness of the proposed approach, we perform experiments on three datasets, namely ARL-VTF, SF-TL54 and RWTH-Aachen. All contain thermal face samples from different variations ranging from frontal faces to faces poses and expressions.

The large *ARL-Visible Thermal Face* dataset [16] (ARL-VTF) contains a collection of paired visible and thermal face images with a spatial resolution of 640×512 from 395 subjects, with over 500,000 images including *baseline* (frontal faces), *occlusion* (eyeglasses), *expression* (lips movements) and *pose* (yaw angles beyond $\pm 20^\circ$) sequences. Face images were acquired in a controlled environment (indoor) at a distance of 2.1m in a time synchronized manner and included *eye, nose* and *mouth* key points annotations. Following the established evaluation protocol, $131,583$ images from 295 subjects were used for training and $5,590$ images from 100 subjects were used for testing.

The *Speaking Faces - Thermal Landmarks 54* dataset [10] (SF-TL54) includes thermal faces with their visible

[2] $F^f_{w_c\times h_c,n}$ encompasses the points $\{(x^f_{up}, y^f_{up}), (x^f_{down}, y^f_{down})\}$, denoting the upper left and bottom right coordinates corner of the f-th face bounding box, respectively, further marked by the plotted points on the Figure 4.

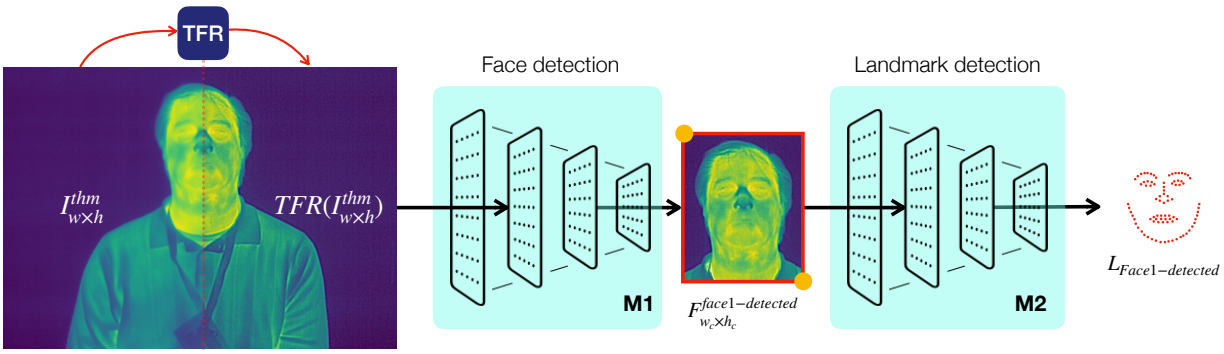

Figure 4. **Illustration of the TFLD pipeline.** A TFR filter is first applied to a thermal image $I^{thm}_{w\times h}$. Hence, the network is fed by an enhanced $TFR(I^{thm}_{w\times h})$ thermal image, where $M1$ is responsible of the face detection $F^{face1-detected}_{w_c\times h_c}$, whereas $M2$ is dedicated to extract a set of facial landmarks $L_{face1-detected}$.

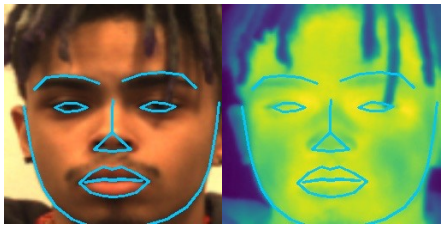

Figure 5. **Landmarks transfer.** Given a synchronized *visible-thermal* paired face and a post-alignment processing, facial landmarks are extracted from the visible face (left) and transferred to the thermal counterpart face (right).

face counterparts of 142 different identities, with a spatial resolution of 464×348. The acquisition was conducted in two stages, where subjects have been captured with a neutral attitude first, then were asked to read a series of short texts. A total number of $2,556$ images have been collected under 9 angles, combining yaw and pitch rotations, in a controlled environment (indoor) along with 54 annotated landmarks. This dataset therefore distinguishes itself with many non-frontal faces, overshadowing the expression deformations which remain subtle. We thus denote this database as *Pose sequences* in our experiments. Following the established protocol, 100 subjects are used for training and the remaining subjects were used for testing.

The *RWTH-Aachen* [8] dataset offers the highest spatial resolution with 1024×768 images with full manual annotations of 68 facial landmarks. It comprises 90 subjects, showing posed expressions, where $2,403$ images and 532 images were used for the training and testing sets, respectively. We thus denote this database as *Expression sequences*.

4.1.2 Evaluation metrics

TFLD is based on face and subsequent landmark detection. Therefore we present in this section the metrics used for quantitatively evaluating (i) thermal face detection and (ii) thermal landmark detection.

Face detection - Model $M1$

Face detection capacity is evaluated by the *Detection Rate* (DR) metric. According to Equation (9) and given a thermal image $I^{thm}_{w\times h,n}$ containing F_n faces, the cardinality defined as the number of faces properly detected by the model $M1$ is expressed as

$$| M1(I^{thm}_{w\times h,n}) | = F. \tag{12}$$

F denotes the number of correctly detected faces. A detected face $F^{f}_{w_c\times h_c,n}$ is only considered as correctly detected, in case that the sub-cropping image of size $w_c \times h_c$ contains at least the eyes and mouth ground truth annotations. Therefore, the face DR per image is given by the ratio $|M1(I^{thm}_{w\times h,n})|/F_n$ and the total face DR assessed by $M1$, noted DR_{M1} is formulated as follows.

$$DR_{M1} = \frac{1}{N}\sum_{n=1}^{N}\frac{| M1(I^{thm}_{w\times h,n}) |}{F_n}, \tag{13}$$

where N represents the total number of images tested, and F_n the number of faces that are annotated in the n-th image.

Landmark detection - Model $M2$

The localization performance is evaluated by the *Normalized Point-to-Point Error* (NPPE) metric and the *Normalized Mean Error* (NME). Given a set of N testing thermal images $\{I^{thm}_{w\times h,n}\}^{N}_{n=1}$ and a detected $F^{f}_{w_c\times h_c,n}$ face, the NPPE metric computed for a particular landmark $k \in [1, K]$ in the f-th face present in the n-th image is referred to as $P^{f,n}_{k}$ and defined as follows.

$$P^{f,n}_{k} = \frac{\|l^{f,n}_{k} - \hat{l}^{f,n}_{k}\|}{d_{IOD}}, \tag{14}$$

where l is the ground truth coordinate and $\hat{l}$ the estimated coordinate provided by $M2$. The quantity d_{IOD} represents the inter-ocular distance (considering the outer corners of the eyes). The NME metric is further used to assess the average performance and is obtained by

$$NME = \sum_{f,k}^{\mathcal{F},K} \frac{P_k^{f,n}}{\mathcal{F} \times K}, \tag{15}$$

where $\mathcal{F}$ indicates the total number of faces in the set.

TFLD aims to provide a higher DR obtained in Equation (13) while a lower NME expressed in Equation (15).

4.2. Performances

To evaluate the effectiveness of the TFLD approaches, we conduct a series of tests on the ARL-VTF, SF-TL54 and RWTH-Aachen datasets including *baseline*, *expression* and *pose* sequences under several variations: *Raw*[3], *Occlusion* and *Poor image quality* along with (w/) or without (w/o) applying the TFR filter.

4.2.1 Face and Landmark detection

Model $M1$ is responsible for face detection and demonstrates a perfect detection rate DR_{M1} as 100% faces are detected in all datasets, under raw, occlusion or low resolution variations, with or without applying the TFR filter.

Model $M2$ locates facial key points and has been trained on each dataset separately, as the datasets differ in amount and semantic definition of facial landmarks. Figures 6, 7, and 8 visualize the landmarks estimated by TFLD on the ARL-VTF, SF-TL54 and RWTH-Aachen datasets, respectively. We observe that facial key points are detected accurately on baseline, expression and pose sequences. In order to highlight the TFR benefits for TFLD, Figure 6 displays *Raw* and *Poor Resolution* samples against *Sharp* samples obtained via TFR. However, detection is challenged in all sequences of poor image quality and occlusion. Nevertheless, we observe that occlusion related to glasses (see Figure 7) does not seem to impact the correct identification of facial key points.

Table 1 illustrates quantitative results w.r.t. NME. Note that landmarks are predicted on images where prior faces are detected by $M1$. The reported performances are stable across datasets. TFR improves performance consistently for all settings, however occlusion and low resolution have a minor impact on performance.

[3]Raw denotes the original image acquisition emerging from the thermal sensor.

4.2.2 Comparison with State-of-the-Art

Table 2 summarizes the performance of landmark detection w.r.t. NME scores, in accordance to State-of-the-Art on the SF-TL54 and RWTH-Aachen datasets, respectively. We note that to the best of our knowledge, we are the first to present results on the ARL-VTF dataset, see Table 1 under the *Raw* setting with TFR processing. TFLD achieves promising results compared to State-of-the-Art methods based on machine learning or deep learning.

4.3. Impact of face alignment on CFR

We proceed here to demonstrate that automated face alignment, as a pre-processing step, is beneficial for Cross-spectral face recognition (CFR). In particular, an appropriate face alignment method is instrumental in improving matching scores. We here evaluate the impact of face alignment in thermal images w.r.t. CFR by comparing FR scores originating from method (i) cropping facial images following bounding box detection, (ii) aligned facial image based on facial key points.

Cropping face images based on the face bounding box consists of cropping the detected face provided by $M1$ to a target size, whereas face alignment based on key points is the result of affine transformations such as *translation, scale* and *rotation*, in order to canonically align the face as [1] with the geometric eye center[4], nose and mouth corners.

Finally, matching is performed using the ArcFace [5] matcher. Table 3 summarizes facial recognition scores, reporting the Area Under the Curve (AUC) metric, computed between the visible gallery face and the associated aligned thermal probe face. A higher AUC indicates a better performance. Note that RWTH-Aachen cannot be considered in this experiment, as it does not include paired visible-thermal faces.

The first naive observation has to do with the low performance of face matching, given that faces are provided without alignment. However, alignment based on the above five key points improves the scores. In particular, when comparing AUC scores of images aligned with ground truth annotations, Dlib (optimizer) [10] decreases the score, while our TFLD method slightly improves the score. Therefore, beside being operational in a wide range of adversarial conditions, TFLD demonstrates its potential as a robust and accurate *landmark annotator* and instrumental in CFR systems.

4.4. Discussion

Detection in a two-stage process, namely *face* and *landmarks*, allows for reliable multi-face detection and landmark detection in unconstrained settings. In particular, this sequential detection prevents the detection of false landmarks where no face appears. Figures 1, 9 and 10 illustrate

[4]The geometric eye center is semantically defined as the midpoint between the outer corners of the eye.

Table 1. **Landmark detection performance** represented by the Normalized Mean Error (NME), on ARL-VTF, SF-TL54 and RWTH-Aachen datasets.

TFLD - $M2$	**ARL-VTF [16]**			**SF-TL54 [10]**	**RWTH-Aachen [8]**
Landmark detection - NME	*Baseline*	*Expression*	*Pose*	*Pose*	*Expression*
Raw *w/o* TFR	0.05244	0.05468	0.06804	0.03082	0.03311
Raw *w/* TFR	**0.05201**	**0.05460**	**0.06737**	**0.03001**	**0.03289**
Occlusion *w/o* TFR	0.05958	0.07328	0.07811	0.04243	0.03975
Occlusion *w/* TFR	0.05549	0.06170	0.07789	0.03856	0.03897
Low Resolution *w/o* TFR	0.07099	0.08861	0.09378	0.05205	0.04167
Low Resolution *w/* TFR	0.06934	0.08605	0.09118	0.04999	0.04039

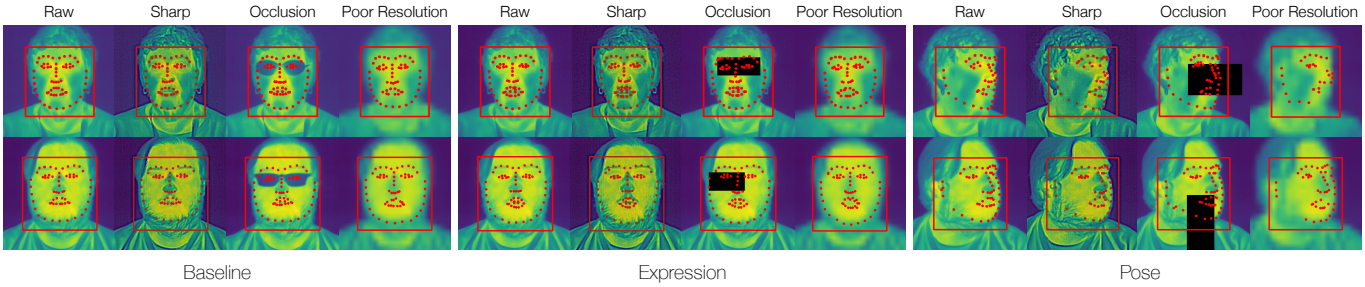

Figure 6. **Visualized faces and landmarks as detected by TFLD**. The face is first detected (red box) followed by landmark detection (red points). TFLD is challenged with *Baseline*, *Expression* and *Pose* sequences, comparing *Raw*, *Sharp (TFR)*, *Occlusion* and *Poor resolution* degradations.

Table 2. **NME score comparison** of TFLD with other approaches on different dataset.

Authors	Methods	SF-TL54	RWTH-Aachen
Mallat *et al.*[11]	AAM	-	0.143
	DAN	-	0.146
Chu *et al.* [4]	Dlib (adapted)	-	0.095
	U-net (multitask)	-	0.040
Kuzdeuov *et al.* [10]	Dlib (optimizer)	0.033	0.057
	U-net	0.035	0.058
Ours	TFLD	**0.03001**	**0.03289**

Table 3. **Evaluation of the impact of face alignment** in thermal images toward a cross-spectral face recognition system with respect to face recognition matching scores AUC %.

Alignment based on	**ARL-VTF**	**SF-TL54**
Bounding box (no alignment)	52.86	60.89
GT annotations	56.03	69.20
Dlib (optimizer) [10] annotations	55.56	67.84
TFLD annotations	**56.93**	**69.59**

a real world scenario, where subjects are captured at different distances, in indoor as well as outdoor environments. In such settings, image quality decreases drastically, impeding localization of facial features. Data augmentation addresses such poor image quality issues and the TFLD model is able to successfully detect face and landmarks due to the enhancement of the TFR filter.

Facial annotations are expensive and tedious to obtain. The ARL-VTF dataset enabled visible-to-thermal landmark transfer, allowing for an increase of number of landmarks fed to TFLD for training. By taking advantage of having synchronized images, our pipeline can circumvent this issue. The infusion of additional landmarks has a significant impact on improving e.g., face recognition performances. To the best of our knowledge, few datasets comply with these characteristics, thus reducing on the one hand the possibility of creating a more unified method of landmark detection (even spread over several databases), and on the other hand limiting the possibility of considering them as benchmarks for face and landmark detection.

5. Conclusions

Accurate and reliable *automatic face and landmark detection* is a key preprocessing step in cross-spectral face recognition. In this paper, we have proposed a novel thermal face and landmark detector (TFLD) that accurately localizes faces and landmarks in the wild. The proposed TFLD sequentially detects face and landmarks, thereby improving the accuracy of landmark localization. Experiments on three datasets suggest that our proposed TFLD achieves competitive results, even under pose and expression variations. The model has additionally been tested on unconstrained thermal images. Finally, we have demonstrated the positive impact of face alignment based on TFLD on cross-spectral face recognition. Future work will involve the detection of additional salient facial features, detected for the purpose of robust cross-spectral face recognition.

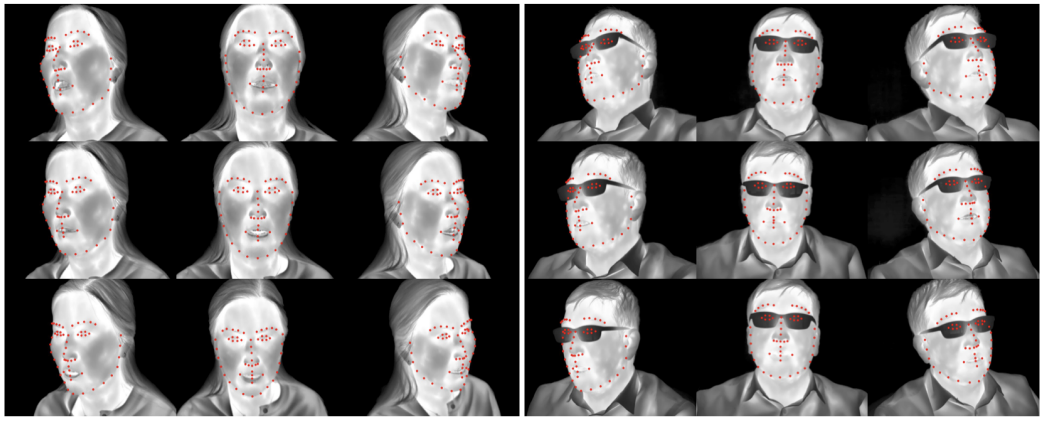

Figure 7. **Visualization of the landmark detection performed by TFLD model on the SF-TL54 dataset [10].** TFLD appears robust to *Pose* variations.

Figure 8. **Visualization of the landmark detection performed by TFLD model on the RWTH-Aachen dataset [8].** TFLD appears robust to *Expression* variations.

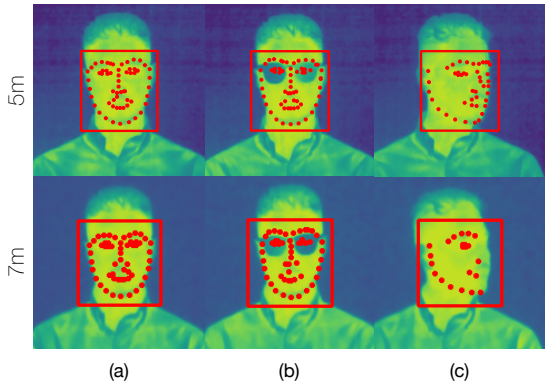

Figure 9. **Examples of TFLD in unconstrained thermal images.** The top row shows images acquired at an offset distance of 5m, whereas the bottom row - at 7m including the co-variates (a) frontal pose variation, (b) eye glasses, and (c) face in profile.

Figure 10. **Examples of TFLD operating in an outdoor environment with sunny weather.** Despite challenging atmospheric conditions, faces, eyes, eyesbrows, nose, mouth and jawline key points are successfully located by our model. (image from TFW [9] dataset).

References

[1] D. Anghelone, C. Chen, P. Faure, A. Ross, and A. Dantcheva. Explainable thermal to visible face recognition using latent-guided generative adversarial network. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[2] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva. Beyond the visible: A survey on cross-spectral face recognition. *arXiv preprint arXiv:2201.04435*, 2022.

[3] L. Assirati, N. R. d. Silva, L. Berton, A. d. A. Lopes, and O. M. Bruno. Performing edge detection by difference of gaussians using q-gaussian kernels. In *Journal of Physics: Conference Series*, volume 490, pages 012–020. IOP Publishing, 2014.

[4] W.-T. Chu and Y.-H. Liu. Thermal facial landmark detection by deep multi-task learning. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[6] A. K. Jain, D. Deb, and J. J. Engelsma. Biometrics: Trust, but verify. *to appear in IEEE Trans. Biometrics, Behavior and Identity Science*, 2021.

[7] J. Keong, X. Dong, Z. Jin, K. Mallat, and J.-L. Dugelay. Multi-spectral facial landmark detection. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020.

[8] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof. A thermal infrared face database with facial landmarks and emotion labels. *IEEE Transactions on Instrumentation and Measurement*, 2018.

[9] A. Kuzdeuov, D. Aubakirova, D. Koishigarina, and H. A. Varol. Tfw: Annotated thermal faces in the wild dataset. 2022.

[10] A. Kuzdeuov, D. Koishigarina, D. Aubakirova, S. Abushakimova, and H. A. Varol. Sf-tl54: A thermal facial landmark dataset with visual pairs. In *2022 IEEE/SICE International Symposium on System Integration (SII)*, pages 748–753. IEEE, 2022.

[11] K. Mallat and J.-L. Dugelay. Facial landmark detection on thermal data via fully annotated visible-to-thermal data synthesis. In *IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020.

[12] K. Nagumo, T. Kobayashi, K. Oiwa, and A. Nozawa. Face alignment in thermal infrared images using cascaded shape regression. *International Journal of Environmental Research and Public Health*, 18(4):1776, 2021.

[13] N. Peri, J. Gleason, C. D. Castillo, T. Bourlai, V. M. Patel, and R. Chellappa. A synthesis-based approach for thermal-to-visible face verification. In *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021.

[14] D. Poster, S. Hu, N. Nasrabadi, and B. Riggan. An examination of deep-learning based landmark detection methods on thermal face imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[15] D. Poster, S. Hu, N. J. Short, B. Riggan, and N. Nasrabadi. Visible-to-thermal transfer learning for facial landmark detection. *IEEE Access*, 2021.

[16] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi, et al. A large-scale, time-synchronized visible and thermal face dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1559–1568, 2021.

[17] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.