The Use of AI for Thermal Emotion Recognition: A Review of Problems and Limitations in Standard Design and Data

Catherine Ordun
Department of Information Systems
University of Maryland
Baltimore County
Booz Allen Hamilton
cordun1@umbc.edu

Edward Raff
Booz Allen Hamilton
Department of Computer Science
University of Maryland
Baltimore County
raff_edward@bah.com

Sanjay Purushotham
Department of Information Systems
University of Maryland
Baltimore County
psanjay@umbc.edu

Abstract

With the increased attention on thermal imagery for Covid-19 screening, the public sector may believe there are new opportunities to exploit thermal as a modality for computer vision and AI. However, thermal physiology research has been ongoing since the late nineties. This research lies at the intersections of medicine, psychology, machine learning, optics, and affective computing. We will review the known factors of thermal vs. RGB imaging for facial emotion recognition. But we also propose that thermal imagery may provide a semi-anonymous modality for computer vision, over RGB, which has been plagued by misuse in facial recognition. However, the transition to adopting thermal imagery as a source for any human-centered AI task is not easy and relies on the availability of high fidelity data sources across multiple demographics and thorough validation. This paper takes the reader on a short review of machine learning in thermal FER and the limitations of collecting and developing thermal FER data for AI training. Our motivation is to provide an introductory overview into recent advances for thermal FER and stimulate conversation about the limitations in current datasets.

Introduction

Computer vision algorithms that use data from the visible spectrum (e.g. RGB) face a variety of challenges when it comes to human Facial Emotion Recognition (FER) due to the representation of superficial facial features laying on the epidermis. Physiological response from stress, fatigue, or other stimuli cannot be visualized on RGB but can be visualized through thermal imagery due to the changes in temperature detected sub-cutaneously. Thermal image data that can capture temperature changes correlated to human vital signs can be a powerful set of data for telemedicine applications supporting healthcare providers as a diagnostic tool for assessing inflammation and stress (Kosonogov et al. 2017). Skin temperature can correlate to certain vital signs and offers a non-invasive method to remotely assess patients. As the cost of high resolution thermal sensors decline and more researchers release thermal FER datasets, there is a great potential to apply thermal imagery for telemedicine purposes. Since the Covid-19 pandemic, governments around the world have begun using thermal sensors combined with AI tools for Covid temperature screening (Ting et al. 2020). From the U.K, China, Italy, Australia, to the U.S., multiple companies are offering the promise of integrated thermal sensing with facial recognition (FR) (Van Natta et al. 2020). We believe that with broader adoption of thermal FR due to changes in HIPAA rules due to Covid-19, it will only be natural that researchers will want to advance their technology towards emotion screening. We caution that before leaping to thermal FER, researchers should be fully aware of the restrictions and limitations of thermal imagery and the problems that may underlie existing thermal FER databases. The adoption of thermal imagery as a source for any human-centered AI task is not easy. Thus, the goal of this paper is to present the state of the literature and discuss the challenges hindering the full adoption of AI as a tool for thermal FER.

Figure 1: RGB, near infrared and thermal images of a resting (up) and fatigued (down) face. In the thermal images, darker pixels corresponds to colder and lighter to hotter. (Lopez, del Blanco, and Garcia 2017)

Advantages of Thermal over Visible

When the public sector thinks about FER and facial recognition (FR), the go-to modality is the visible spectrum usually encoded as RGB. RGB images have dominated the area of FER, indicative through a variety of well known facial

databases used in AI.[1] But, FR using RGB databases has become a controversial area of computer science, requiring careful consideration of its flaws and innate assumptions within the data (Martinez-Martin 2019; Buolamwini and Gebru 2018; Greene, Hoffmann, and Stark 2019; Singer and Metz 2019; Lohr 2018). Beyond the original intended academic purposes, some RGB databases have been taken down in order to prevent industry FR training (Murgia 2019). In the wake of Black Lives Matters protests in June 2020, Microsoft and IBM discontinued their development of FR, where Amazon invoked a one year moratorium on FR based on evidence of algorithmic discrimination against communities of color (Matsakis 2020). Of particular value to the public sector, is whether thermal imagery for FER affords any level of privacy protection and bias mitigation. The answer may stem from the separation of thermal imagery from other machine learning tasks, known to increase recognition and decrease anonymity.

Figure 2: Example of data from the Iris dataset (Hammoud)

We believe that long-wave Infrared Radiation (LWIR) used alone, as a data source for FER, may be able to provide some form of anonymity for healthcare applications to minimize racial, ethnic, and potentially gender bias, when compared to RGB for FER. Through its low, grey-scale resolution [2] and reliance on temperature vectors driven by underlying vasculature (Ioannou, Gallese, and Merla 2014), rather than superficial skin tone, texture, and pigmentation, thermal imagery can be more challenging to easily identify individuals. But there still remains a variety of issues to preserve privacy. For example, anonymity may not be possible if thermal FER is combined with the machine learning task of FR, especially since thermal FR is well researched with multiple methods proposed to detect and recognize individuals. The concept of separating FR from other tasks is not uncommon. Ethicists Van Natta, et al. (Van Natta et al. 2020), question whether during Covid-19 temperature monitoring, there is even a need to conduct FR given how the overall purpose is to identify infection as opposed to identity. It is important to caution, that although thermal FR is more challenging than the visible domain, it is feasible to use thermal imagery as a "soft" biometric due to its invariance under lighting and pose (Reid et al. 2013; Friedrich and Yeshurun 2002). For example, superficial vascular networks are unique to each person's face as proposed by Buddharaju et al. (Buddharaju et al. 2007), and can be extracted through methods like anisotropic diffusion to identify minutiae points akin to fingerprints as shown in Figure 3. Further, combining RGB with thermal can increase recognition accuracy. For example, Nguyen et al. (Nguyen and Park 2016) used a combination of thermal and visible full body images for gender detection, finding that their proposed method of score-level fusion (training two separate SVM classifiers) combining thermal and visible led to a decrease in error of 14.672 equal error rate (EER) when compared to using thermal only (19.583 EER) and visible only (16.540 EER).

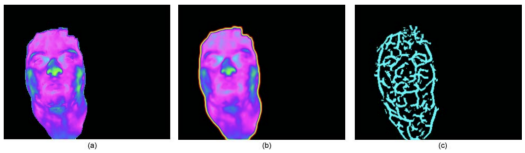

Figure 3: Vascular network extraction: (a) Original segmented image; (b) Anisotropically diffused image; (c) Blood vessels extracted using white top hat segmentation, per (Buddharaju et al. 2007)

In addition, there has been research in the computer and electrical engineering fields to develop sensor-level privacy for thermal sensors in situations where people need to be sensed and tracked, but not identified. Work by Pittaluga et al. (Pittaluga, Zivkovic, and Koppal 2016) demonstrated different techniques to include digitization that masks human temperatures measurements thereby obscuring any ability to detect faces shown in Figure 4, manipulating the sensor noise parameters as the thermal image is being generated, and algorithms to under or overexpose specific pixels that are designated as "no capture" zones. Still in research, these techniques require different levels of hardware and firmware upgrades based on the thermal sensor.

Thermal imagery has additional technical advantages including how it is (1) invariant to lighting conditions unlike RGB, allowing the detection of physiological response (heat) to occur in low light or total darkness; (2) is a reliable and accurate correlation to standard physiological measures like respiration and heart rate; (3) is non-invasive i.e., requiring no skin contact whatsoever, making it convenient and non-intrusive and potentially relevant for non-communicative persons; (4) resistant to intentional deceit since physiological responses cannot be faked, whereas visible facial expressions can be controlled; and (5) is able to reveal facial disguises (i.e. wigs, masks) since these materials have high reflectivity and display as the brightest on thermograms compared to human skin which is among the

[1] CK+ , FER 2013 , FERET , EmotioNet, RECOLA, Affectiva-MIT Facial Expression Dataset, NovaEmotions, MultiPIE, McMaster Shoulder Pain, AffectNet, Aff-Wild2, the Japanese Female Facial Facial Expression database, and CASME II for microexpressions

[2] Thermal imaging manufacturers offer a variety of color palettes for visualizing temperature beyond "white hot" such as "iron bow" and "rainbow". It should be cautioned that some manufacturers offer fusion visualizations that fuse the RGB and thermal images together thereby improving resolution.

Figure 4: Digitization privacy in different scenes: digitization results in scenes with people, computers and buildings. The left column are the input 16 bit images and the right column is the simulated output. (Pittaluga, Zivkovic, and Koppal 2016)

darkest objects with low reflectivity (Pavlidis and Symosek 2000). In addition, thermal imagery offers physiological signals of social interactions from person to person. In terms of deceit detection, it is valuable to note that RGB images can also be used to detect microexpressions using databases like CASME II. Microexpressions are genuine, quick facial movements that may be uncontrollable or unnoticeable by the individual, and therefore have been studied as an indication of deception (Yan et al. 2014). The RGB images used for studying microexpressions, however, are different than standard RGB FR datasets. They consist of video sequences captured using spontaneous natural elicitation, captured at a high frame rate of 200 fps, and labeled with facial action units (FAUs) which are encoded combination of facial movements based on Paul Ekman's Facial Action Coding System (FACS) (Ekman 1999).

Physiology and Thermal FER

A brief explanation of thermal radiation helps to understand how facial skin acts as a radiating surface. Thermal radiation is emitted by all objects above absolute zero (-273.15 °C). Human skin is estimated at 0.98 to 0.99 ϵ (Yoshitomi et al. 2000). The principal of thermal image generation is well understood by the Stefan-Boltzmann law that states total emitted radiation over time by a black body is proportional to T^4 where T is temperature in Kelvins: $W = \epsilon\sigma T^4$ where W is radiant emittance (W/cm^2), ϵ is emissivity, σ is the Stefan-Boltzmann constant ($5.6705 \cdot 10^{-12} W/cm^2K^4$), and T is Temperature (K).

A black body is an object that absorbs all electromagnetic radiation it comes in contact with. No electromagnetic radiation passes through the black body and none is reflected. Since no visible light is reflected or transmitted, the object

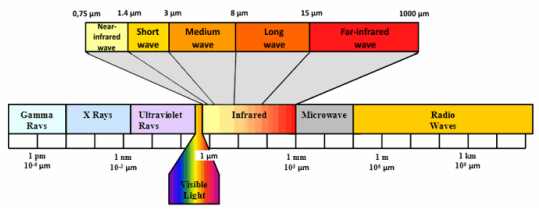

Figure 5: Long-Wave IR falls in the wavelength range of $8\,\mu m$ to $15\,\mu m$

looks black upon visualization from thermal imagery, when it is cold. Thermal sensors respond to infrared radiation (IR) and produce visualizations of surface temperature. Because LWIR operates in a sub-band of the electromagnetic spectrum per Figure 5 it is invariant to illuminating conditions meaning that it can operate in low light to complete darkness. By imaging temperature variations to emotionally induced stimuli such as videos or pictures, thermograms reveal genuine responses to social situations. This occurs through activation of the autonomic nervous system (ANS) where emotional arousal leads to a perfusion of blood vessels innervated at the surface of the skin (Ioannou, Gallese, and Merla 2014).

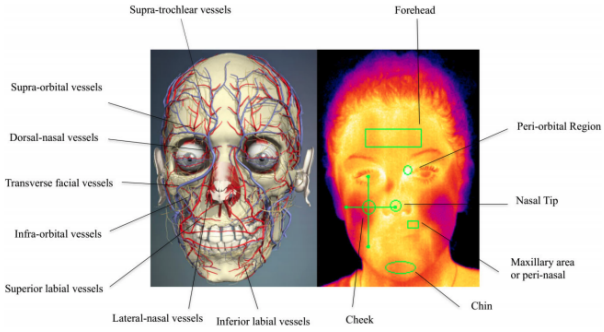

Figure 6: Thermal representation for extraction of ROIs by Ioannou

These images are called thermograms and are the data captured in thermal FER datasets, with labels based of the emotional response elicited (i.e. happiness, disgust, sadness, deceit, stress, etc.). Although today's need for a touch-less system are paramount, the concept of using thermograms for contact-less physiological monitoring is not new and rooted in the intersection of physiological research (Selinger 2016;Buddharaju 2007;Pavlidis 2000; Ionnou 2014) and affective computing (Wilder 1996;Yoshitomi 2000;Goulart 2019). These include applications for FER where different emotions are detected from thermal facial images alone, in addition to person re-identification on thermal imagery, for FR. Since 1996 (Wilder et al. 1996) there have been numerous studies evaluating how thermograms correlate with vital measures. In 2007, Pavlidis (Pavlidis et al. 2007) demonstrated that thermal imagery is a reliable measure to assess emotional arousal where different regions of the face (zygo-

Table 1: Thermal Facial Emotion Recognition Datasets

Dataset	Year	Pose	Pairs	Affect	Subj	Access	Seq	Multi	THR	VIS
Univ. Notre Dame (UND)	2002	Spont.	Yes	UNK	241	R	UNK	Yes	LWIR	Yes
Equinox (Equinox ; Heo et al. 2004)	2004	Posed	UNK	3	90	N/A	No	No	MW, LWIR	Yes
IIT Delhi (Kumar)	2007	Posed	UNK	UNK	108	R	No	UNK	NIR	No
Univ. Houston (Buddharaju et al. 2007)	2007	Both	Yes	0	138	UNK	UNK	No	MWIR	Yes
SC-Face (Grgic)	2009	None	Yes	0	130	R	No	No	NIR	Yes
USTC-NVIE (Wang et al. 2010)	2010	Both	UNK	6	100	N/A	Yes	No	LWIR	Yes
Zhang (Zhang et al. 2010)	2010	Posed	UNK	0	350	R	No	UNK	NIR	No
UCHThermalFace (Hermosilla et al. 2012)	2012	Posed	No	3	102	UNK	Yes	No	LWIR	UNK
KTFE Database (Nguyen et al. 2013)	2013	Spont.	Yes	7	26	UNK	Yes	No	LWIR	Yes
Iris (Hammoud)	2013	Posed	Yes	3	30	P	No	No	LWIR	Yes
RGB-D-T (Simón et al. 2016)	2016	Posed	Yes	5	51	UNK	UNK	UNK	LWIR	Yes
VIS-TH (Eurecom) (Mallat and Dugelay 2018)	2018	Posed	Yes	4	50	R	Yes	Yes	LWIR	Yes
RWTH Aachen Univ. (Kopaczka, Kolk, and Merhof 2018)	2018	Posed	No	8	90	R	Yes	UNK	LWIR	No
Tufts Face Database (Panetta et al. 2018)	2018	Posed	Yes	5	113	R	Yes	No	NIR, LWIR	Yes
UL-FMTV (Ghiass et al. 2014)	2018	Posed	Yes	UNK	238	R	Yes	Yes	N, MW, LWIR	No
ThermalWorld (Kniaz et al. 2018)	2019	Spont.	Yes	0	516	R	No	No	LWIR	Yes
RFLDDJ (Seo and Chung 2019)	2019	UNK	Yes	UNK	UNK	P	UNK	No	LWIR	Yes

Dataset - Database name, Year - publication year, Pose - Posed, Spontaneous, or Both, Pairs - Visible and Thermal, Affect - Number of labeled expressions, Subj - Number of unique human subjects, Access - R (requires permission from authors), P (publicly downloadable), Seq - Yes or No for availability in dataset of video sequences, Multi- Yes or No for multi-session recording, THR - Thermal image modality, VIS - Yes or No for presence of visible images, UNK means information was not provided in the paper.

maticus, frontal, orbital, buccal, oral, nasal) correlate with different emotional responses. Thermal imagery also visualizes the physiology of perspiration (Pavlidis et al. 2012; Ebisch et al. 2012), cutaneous and subcutaneous temperature variations (Hahn et al. 2012; Merla et al. 2004), blood flow (Puri et al. 2005), cardiac pulse (Garbey et al. 2007), and metabolic breathing patterns (Pavlidis et al. 2012) and has been used to monitor heat stress and exertion (Bourlai et al. 2012). The reliability of thermal temperature readings have been repeatedly shown to be consistent and correlate accurately with gold standard physiological measures of electrocardiography (ECG), piezoelectric thorax stripe for breathing monitoring, nasal thermistors, skin conductance, or galvanic skin response (GSR) (Pavlidis et al. 2007; Sonkusare et al. 2019).

We can even observe these changes with the naked eye, such as embarrassment causing a person to blush (Sonkusare et al. 2019), or fear leading to pallor (Kosonogov et al. 2017). Merla (Merla 2014) offered a survey of thermal studies in psychophysiology from 1990 to 2013, demonstrating a series of emotional responses detected on thermal imagery such as startle response, fear of pain, lie detection, mental workload, empathy, and guilt. These responses occur in different regions of the face, or ROIs. Salazar-Lopez found high arousal images elicited temperature increases on the tip of the nose (Salazar-López et al. 2015). Kosnogov (Kosonogov et al. 2017) found that more arousing an image, the faster and greater the thermal response on the tip of the nose. He speculated that the speed and magnitude of these thermal responses were linked to autonomic adjustments normal to emotional situations. Zhu (Zhu, Tsiamyrtzis, and Pavlidis 2007) found that deception was detected through increased forehead temperature and Puri (Puri et al. 2005) found the forehead to be correlated with stress. Social responses based on one-on-one personal contact can also be observed. For example, Ebisch (Ebisch et al. 2012) found "affective synchronization" of facial thermal responses between mother and child, where distress temperatures at the tip of the nose were mimicked by the mother as she watched her child in distress. Fernandez (Fernández-Cuevas et al. 2015) summarizes analysis by Ioannou et al. (Ioannou, Gallese, and Merla 2014) describing whether temperature increases, decreases, or stays the same based on different emotions and ROIs provided in Figure 7.

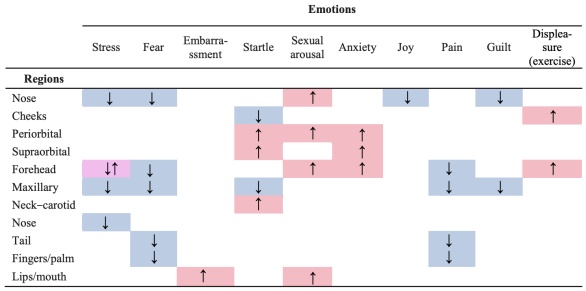

Figure 7: Skin thermal variations in the considered regions of interest across emotions

AI and Thermal FER

Since 2000 with (Yoshitomi et al. 2000), machine learning in thermal FER has grown slowly to include emotion classification by (Khan, Ingleby, and Ward 2006; Nhan and Chau 2009; Wang et al. 2014a; Jarlier et al. 2011; Wang et al. 2014b; Trujillo et al. 2005) with gradual adoption of AI methods such as neural networks. The ability to move away from manual, hand-crafted feature extraction to automatic learning through neural networks has already proven advantageous for thermal-to-visible image translation through GANs (Mallat et al. 2019; Kniaz et al. 2018; Chen and Ross 2019), and for automated temperature vector extraction of facial ROIs (Sonkusare et al. 2019). Earlier works in deep learning applied to thermal FER such as the works of Wang et al. in 2014 (Wang et al. 2014a) using a

Table 2: Selected Thermal Facial Emotion Recognition AI Papers

Author	Year	Affect	ROIs	Model	Dataset	Target	Acc	Data	Code	Params
Stemberger	2010	Cognitive Workload	7 ROIs	ANN	Custom dataset	Multiple Workload	81.0%	(-)	(-)	(+)
Wang	2014	Spont. Affect	Whole face	DBM	USTC-NVIE	Valence	62.9%	(+)	(-)	(+)
Wu	2016	Posed Affect	Whole face	CNN	RGB-D-T	Multiple Affects	99.40%	(-)	(-)	(-)
Simon	2016	Posed Affect	Whole face	CNN	RGB-D-T	Multiple Affects	UNK	(-)	(-)	(+)
Cho	2017	Stress	Nose	CNN	Custom dataset	Binary Stress	85.59%	(-)	(-)	(+)
Lopez	2017	Exercise Fatigue	Whole face, 3 ROIs	CNN, SVM	Custom dataset	Binary Fatigue	23.3% - 81.8%	(+)	(-)	(+)
Haque	2018	Pain	Whole face	CNN, LSTM	Custom dataset	5 Pain Levels	18.33% CNN	(+)	(-)	(+)
Ilyas	2018	Spont. Affect	Whole face	CNN, LSTM	Custom dataset	Multiple Affects	89.74%	(-)	(-)	(-)
Elbarawy	2019	Posed Affect	Whole face	CNN	Iris	Multiple Affects	96.7%	(+)	(-)	(+)
Ilikci	2019	Posed Affect	Whole face	CNN	Iris	Multiple Affects	92.72%	(+)	(-)	(+)
Shreyas Kamath	2019	Posed Affect	Whole face	CNN	Tufts Face Database	Multiple Affects	96.2%	(+)	(-)	(+)

Year - Publication year, Affect - Expression type (Posed and Spont. mean basic discrete emotions), ROIs - facial regions of interest, Model - Deep learning algorithm type, Dataset - name of database, Target - the predicted class (all papers identified were classification), Acc - Best classification accuracy across models reported. Data - link to database provided if custom or name of public database provided, Code - link to code provided, Params - model parameters disclosed in paper, Annotations of (-) indicate information not disclosed, and (+) means it was disclosed in the paper.

Table 3: Examples of Thermal FER Experimental Design Parameters

Author	Year	Thermal Cam.	Dual Sensor	Thermal Res.	Dem.	Exclusion	Subjects	Temp.	Rest Time	Lighting	Stimulus
Nhan	2010	ThermaCAM	UNK	UNK	9F, 3M, mean 24 yo	UNK	12	UNK	20 min	UNK	Static images
Wang	2010	SAT-HY6850	UNK	320 x 240	58F, 157M, 17 - 31 yo	UNK	215	Means 23.29	UNK	Yes	Emotional videos
Hermosilla	2012	Flir 320 TAU	UNK	324 x 256	UNK	UNK	102	UNK	UNK	UNK	UNK
Nguyen	2013	NEC R300	Yes	UNK	UNK gender, 11 - 32 yo	UNK	26	24 - 26	2 hrs.	UNK	Emotional videos
Salazar-Lopez	2015	ThermoVision A320G	UNK	UNK	60F, 60M, 24 - 27 yo	Yes	120	18 - 25	10 - 15 min.	UNK	Static images
Lopez	2017	Therm-App	UNK	288 x 384	8F, 11M, 23 - 27yo	UNK	19	UNK	Until heart rate below 20 bpm	UNK	Exercise
Mallat	2018	Flir Duo R	Yes	160 x 120	No	UNK	50	25	No	Yes	UNK
Goulart	2019	Therm-App	UNK	384 x 288	8F, 9M, 8 - 12 yo	UNK	17	20 - 24	10 min.	Yes	Questionairre
Sonkusare	2019	Flir A615	UNK	640 x 480	11F, 9 M, 22 - 30 yo	Yes	20	22	No alcohol & caffeine 2 hrs. prior	Yes	Auditory stimulus
Panetta	2020	FLIR Vue Pro	UNK	UNK	UNK	UNK	113	UNK	UNK	Yes	UNK

Year - Publication year, Thermal Cam. - Type of LWIR camera, Dual Sensor - Yes or No, captures visible and thermal simultaneously, Thermal Res. - Reported thermal pixel resolution, Dem. - Demographics of subjects, Exclusion - Yes or No, exclusion or inclusion criteria documented, Subjects - Number of unique human subjects, Temp. - Room temperature for experiment reported in degrees Celsius, Rest Time - Time subjects reach relaxed state prior to image capture, Lighting - Yes or No, illumination design documented, Stimulus - Type of stimulus to provoke spontaneous response, if spontaneous, UNK means information was not found in the paper.

Deep Boltzman Machine (DBM) found that learning feature representations directly from thermal images of the NVIE dataset (Wang et al. 2010) led to greater accuracy (62.9%) when predicting low and high valence, compared to statistical temperature vectors manually extracted from thermal images followed by dimensionality reduction (PCA) and SVM (61.0%). Further, Wang asserted that the DBM learned from features representing a mixture of thermal datasets such as the Equinox (Equinox) and NVIE led to greater accuracy by 5.3%. Thermal features can outperform visible features in FER, overall, even without deep learning methods. Goulart's thermal multi-affect classifier using PCA and LDA outperformed visible emotion classifiers particularly on challenging expressions such as disgust and fear which can range between 40% to 50% for RGB accuracy. Whereas, Goulart's thermal classifiers detected disgust with 89.93% and fear at 88.22% true positive rates (Goulart et al. 2019).

Li et. al. (Li and Deng 2018) describes how AI research in the visible domain grew based on the broad dissemination of public, large-scaled, natural data per Figure 8. Any internet search will reveal dozens of RGB FR databases easily accessible and downloadable, such as the Top 15 list of facial recognition databases on Kaggle (Hamdi 2020). They identified 74 visible "deep FER" papers using CNNs, Generative Adversarial Networks (GANs), Restricted Boltzman Machines (RBM), Deep Auto Encoders, Deep Belief Networks (DBN), and Recurrent Neural Networks (RNN) trained on such RGB FR datasets. But, AI in thermal FER lags behind, possibly due to the lack of large-scale, publicly available, and comprehensive thermal FER datasets.

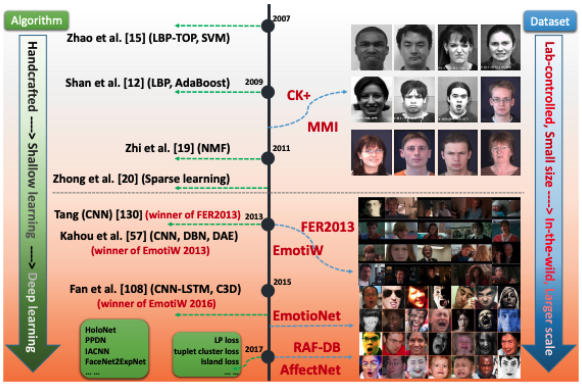

Figure 8: Growth of lab-controlled, small size data to "in-the-wild", larger scale data encouraged use of deep learning algorithms in visible FER (Li and Deng 2018)

Where Li identified 74 papers, we only identified 14 thermal FER datasets in Table 1 whose numbers have increased since 2018 possibly due to the decreasing cost of thermal cameras and the easier ability to purchase them online. Further, we identified only eleven AI thermal FER papers, shown in Table 2, starting in 2010, indicating a slow evolution from manual feature extraction using geometric methods to learning latent representations using deep learning. These works do not consistently release code and have varied levels of explanation around experimental design and arousal stimulus, which we summarized in Table 3. This makes it challenging to reproduce, much less compare across studies. Re-

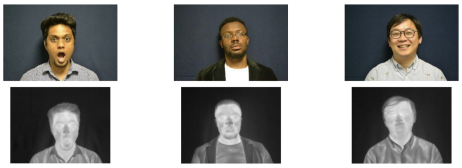

Figure 9: The Tufts Face Database (Panetta et al. 2018)

searchers in thermal emotion recognition such as Goulart et al. (Goulart et al. 2019) agree, particularly since there is no standard thermal FER imaging benchmark dataset consistently used across studies. In an empirical review reproducing 255 machine learning papers, Raff (Raff 2019) notes that papers which are scientifically sound and complete, should be independently reproducible based solely on explanation, details, and descriptions. Failures in reproducibility can occur when language or notation is unclear, when the algorithm is missing details about implementation or equations, and when nuanced details are left out. In Table 1 we catalog the few available (via request or publicly) thermal datasets that have been used for tasks including FR and FER. They vary in scope, where some do not have emotion labels at all, making it difficult to benchmark and standardize results that may eventually impact psychological and health-related decisions. One example of a recently developed thermal FER dataset is by Tufts University shown in Figure 9.

Thermal FER Data Challenges

Some researchers have noticed the lack of variation across thermal FR dataset that fail to account for diverse emotional states, alcohol intake or exercise, and ambient temperature, leading them to doubt the rigor of the reported results especially in real life conditions (Shoja Ghiass 2014). Assuming that the lack of a comprehensive thermal FER benchmark dataset is one factor that hinders the advancement of AI research, we can begin exploring the challenges of designing such a dataset. But, developing a thermal FER dataset is different than simply crawling the web for RGB faces. The collection of thermal FER data requires an experiment unto itself, needing institutional review board (IRB) approval, subject recruitment, experimental design, and specialized equipment. As a result, thermal FER datasets are expensive in terms of time and labor. We have observed some trends across databases that if addressed in the development of a single high-fidelity dataset, may carve a path for greater adoption of thermal AI FER studies. We justify these assertions based on research in the psycho-physiology domain, below.

Include video sequences

Video sequences present timing of the arc of expression onset and delay. It is important to capture intensity and duration of expression which has been found consistent with automatic movement and neuropsychological models (Tian, Kanade, and Cohn 2005). Levenson et al. (Levenson 1988) indicated that duration of an emotional response is 0.5 – 4 seconds. But Nguyen (Nguyen et al. 2013) cites mistakes in many of the leading thermal recognition databases. In the USTC-NVIE database their procedure for data acquisition had video gaps between each emotion clip at 1-2 minutes which is too short for participants to establish a neutral emotion status. Research indicates that for thermal response (cutaneous skin temperature), there is a delay after stimulus that needs to be accounted for and recorded (Ioannou, Gallese, and Merla 2014) and temperature change can occur in less than 30 seconds upon stimulation (Pavlidis et al. 2012). Temperature changes at the tip of the nose can occur as fast as 10 seconds after stimulus and last 20 - 30 seconds regardless of distress or soothing (Ebisch et al. 2012). In a more recent paper, (Sonkusare et al. 2019) were able to quantify the temporal dynamics of thermal response when compared to gold standard measures like Galvanic Skin Response (GSR) demonstrating that thermal response occurred only 2 seconds later than GSR when exposed to an auditory stimulus. Static images without a time axis can be incomplete and will fail to capture the complete physiological signal and emotional response.

Enable spontaneous response

Many existing thermal databases that are focused only on FR have discrete, posed affects based on the labeling defined by Ekman (Ekman 1999). But affective researchers argue that spontaneous emotional reactions are more realistic since, "people show blends of emotional displays. . . hence, the classification of human non-verbal affective feedback into a single "basic"-emotion category may not be realistic." (Gunes and Pantic 2010; McDuff, Girard, and El Kaliouby 2017). Further, multiple emotions typically occur as opposed to a single discrete response. For example, in a 1993 study by Gross et al. 85 subjects self-reported a variety of feelings after watching a close-up arm amputation medical video (Gross and Levenson 1993).

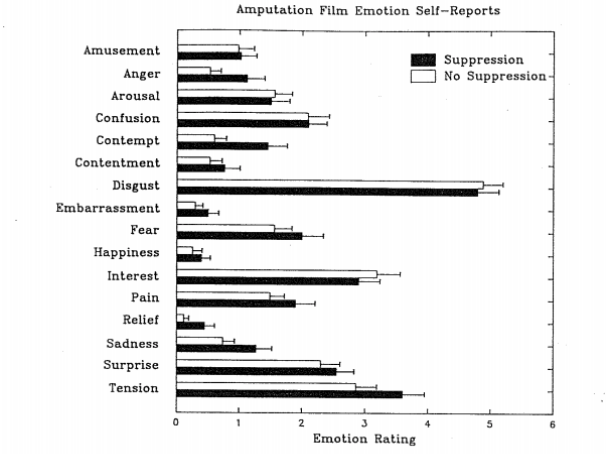

Figure 2. Emotion self-reports by condition for the amputation film, with standard errors of the mean.

Figure 10: Multiple feelings self-reported after exposure to high arousal video (Gross and Levenson 1993)

Another argument against discrete labels is the possibility that people express emotions as internalizers or external-

izers, meaning different people suppress emotional expression in different ways making it difficult to truly capture expression in a basic, discrete manner (Gross and Levenson 1993). To elicit spontaneous response, emotion researchers use static images such as the International Affective Picture System (Kosonogov et al. 2017) or short clips of emotional videos (Nguyen et al. 2013). In a recent 2019 study by Sonkusare et al. (Sonkusare et al. 2019), they use an auditory stimulus described in Figure 11 to mimic a startle response, spontaneously.

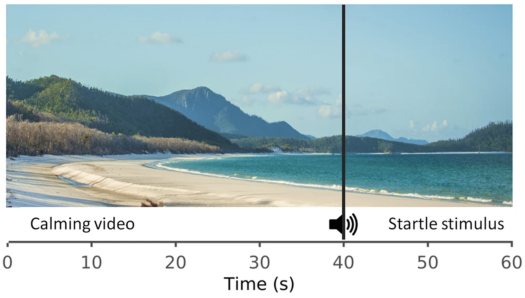

Figure 11: Example of an emotional stimulus by Sonkusare et al. to elicit a spontaneous response. A calming ocean video clip was played for 60 seconds. A loud gunshot sound (80dB) was played at 40seconds to mimic a startle response. (Sonkusare et al. 2019)

Provide social or personal context

In a similar vein to spontaneous, natural emotion collection, providing social context in an experimental setting will change the nature of the emotion recorded. Context labeling to account for elicitation methods that are prompted spontaneously through personal elicitation (i.e. images, videos), versus social interaction with another person (or robot per (Goulart et al. 2019)) may signal different physiological responses reflected in thermal imagery. Factors that influence these responses may include interpersonal distance, gaze direction, and opposite gender in the interaction (Kosonogov et al. 2017; Gunes and Pantic 2010). A sociodynamic model of emotions (Mesquita and Boiger 2014) asserts that emotions "emerge in interplay with and derive their specific function from the social context. This means that emotional experience and behavior will be differently constructed across various contexts". For example, Goulart (Goulart et al. 2019) analyzed emotional response for 17 children during a human-child robot interaction experiment shown in Figure 12. Using Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA), they inferred happiness and surprise as the most frequently expressed, which were consistent with what the children self-reported upon interacting with the New-Mobile Autonomous Robot for Interaction with Autistics (N-MARIA) robot.

Collect multimodal pairs

In 2000 Yoshitomi (Yoshitomi et al. 2000) classified discrete affects by combining visible, thermal, and audio signals from 21 test subjects, achieving 85% accuracy. Zhu et

Figure 12: Experimental setup showing the child-robot interaction by Goulart et al. (Goulart et al. 2019) (a) Before showing the robot; (b) After presenting it.

al. (Zhu, Tsiamyrtzis, and Pavlidis 2007) discussed multimodal data as "cross scale" data for biomedical research, or interconnections of different types of data using AI to infer mappings even if some data is missing. In essence, both were developing multimodal machine learning models, where multiple modalities, or types of information, may be combined to increase the accuracy of models (Baltrušaitis, Ahuja, and Morency 2018). The approach to collect pairs is not new. Nguyen collected thermal FER pairs for the KTFE database (Nguyen et al. 2013) and the Iris (Hammoud), Eurecom (Mallat and Dugelay 2018), and University of Notre Dame (UND) also have pairs which offer greater flexibility for different AI use cases like image translation for person re-identification. This includes research into thermal-to-visible GANs (Mallat and Dugelay 2018; Kniaz et al. 2018; Chen and Ross 2019; Zhang et al. 2018). With paired images capturing the RGB and LWIR images simultaneously using a camera equipped with a dual sensor, offers a mapping between both modalities for an AI algorithm to learn.

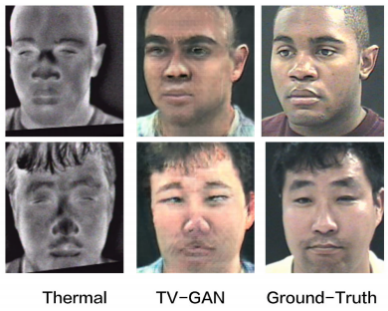

Figure 13: Example of TV-GAN trained on multimodal pairs for thermal-to-visible image translation (Zhang et al. 2018)

Document experimental setup

Documenting experimental setup is important in order to minimize bias in the resulting thermogram, which can be affected by a variety of environmental and human subject conditions. Ioannao (Ioannou, Gallese, and Merla 2014) articulates in his paper on the potential and limitations of thermal imaging in physiology that, "Cutaneous thermal responses to external stimuli of psychophysiological valence could re-

sult in small temperature variations of the ROIs. Thus, it is extremely important to ensure that the observed temperature variations are not artifacts due to either environmental physiological causes or simply subject motion." Some of these can be minimized, the methods of which should be recorded and shared in the paper so that other thermal FER data collection trials can be repeated or improved to control for these external factors.

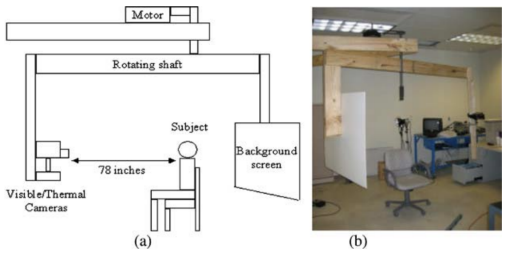

Figure 14: Experimental Setup for Iris dataset capture (Kong et al. 2007)

In Table 3 we provide a sample of experimental parameters from several thermal FER papers and show how they vary from paper to paper. This demonstrates non-standard setups over the years of thermal FER research that could affect the reusability and generalization of these data for AI experiments. But, different papers vary in the extent of how much they document their experimental protocol provided in an example set of papers in Table 3. Multiple factors need to be managed in order to minimize variables in the environment that influence thermal capture, leading to potentially misleading thermograms such as 1) Cold or warm air, as well as humidity, 2) Facial expressions (e.g. open mouth), 3) Physical conditions (e.g. lack of sleep, alcohol, caffeine), 4) Mental state (i.e. fear, stress, excitement), 5) Opaque to glasses, 6) Skin temperature variance through the day (Kosonogov et al. 2017). Fernandez et al. provide a comprehensive review of environmental, individual, and technical factors that influence IR reliability per Figure 15 (Fernández-Cuevas et al. 2015).

Experimental design also includes the demographics of recruited subjects. Very few details are provided about race and ethnicity shown in Table 3 for the exception of (Lopez, del Blanco, and Garcia 2017) who indicated that nine out of 19 individuals were of Chinese ethnicity. With the ethical problems of visible FR in failing to train algorithms on a representative and balanced minority dataset, thermal FER researchers need to understand exactly what subjects are being included in the data and what underlying assumptions are being broadcast into training. Further, we have so far been discussing thermal FER on adults in the various papers introduced. Very few studies, limited to (Goulart et al. 2019) for child-robot interaction, (Ioannou et al. 2013) for guilt, (Ebisch et al. 2012) for child-mother imprinting, (Manini et al. 2013) for mother-child of vicarious autonomic response, collect thermal FER data on children. For the exception of Panetta et al., none of the thermal databases we

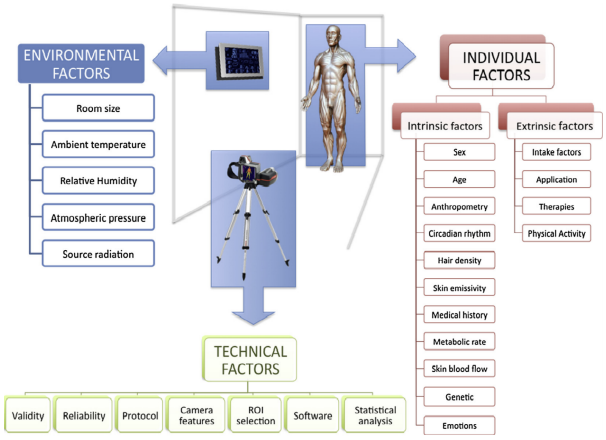

Figure 15: Factors influencing thermal imagery of humans (Fernández-Cuevas et al. 2015)

identified appear to include children in their dataset, to the author's knowledge for thermal FER. So far, much work is still needed to generate an ethnically and age-diverse thermal FER dataset. Lastly, experimental set-up should also document technical methods that aim at normalizing the detected thermal face. For example, Wang et al. (Wang et al. 2014a) describes using the Otsu threshold algorithm to binarize the thermal images, detecting the face boundary, and removing baseline temperature to minimize the effects of temperature changes in the environment. Similar methods were introduced by Friedrich and Yeshurun in 2002 (Friedrich and Yeshurun 2002).

Accounting for Sensor Differences

Lastly, the cost of thermal sensors through vendors like FLIR, have decreased over the past decade with increasingly higher quality resolution made accessible to the public. Prior papers have extensively used the Iris and Equinox (now discontinued) datasets. But with the release of more custom datasets as shown in Table 1, is it fair to compare the output of thermal images from one sensor against another, which may have different optical properties? Or, is it sufficient that each sensor operates in the LWIR band? Many researchers have used different thermal sensors over the years: Pavlidis detected anxiety in thermal imagery in 2000 using an uncooled thermal camera with a spectral band of 8μm-14μm manufactured by Raytheon (the ExplorIR model) (Pavlidis and Symosek 2000), Nguyen in 2014 used a NEC R300 collecting in the 8μm-14μm band (Nguyen et al. 2013), Aureli in 2015 used a FLIR SC660, an uncooled microbolometer sensor that collects in the 7.5μm – 13μm band (Aureli et al. 2015), and Eurecom researchers in 2018 used a FLIR Duo-Pro, an uncooled VOx Microbolometer sensor operating in 7.5μm–13.5 μm (Mallat and Dugelay 2018). Table 3 provides a selection of thermal cameras used across various thermal FER studies as examples of how the cameras vary from study to study.

Table 4: Summary of Thermal FER Data Challenge

Challenge	Consequence	Mitigation	Opportunities
Include video sequences	Static images fail to capture the complete temporal dynamics of emotional response.	Including labeled videos in thermal FER dataset.	Spatio-temporal labeling of thermal onset, delay, duration of physiological response.
Enable spontaneous response	Discrete posed expressions may not invoke realistic physiological response.	Add spontaneous elicitation where possible, in addition to discrete set.	Natural, "in the wild" expressions that offer accurate representations of emotion.
Provide social or personal context	Thermal data collected without social stimuli may not be useable for social use cases.	If appropriate, label social context or if controlling for, document how social response has been minimized.	Social interaction thermal FER expressions, with labeled context and scenarios.
Collect multimodal pairs	No opportunity to increase accuracy or learn from additional modality mappings if only one modality (thermal) is collected.	May require dual sensor, or experimental design for simultaneous capture using two cameras.	Multimodal pairs for various social, spontaneous elicited thermal FER domains.
Document experimental setup	Confounding through uncontrolled environmental variables can lead to misleading images.	Report at minimum, the parameters shown in in Table 3.	Standard thermal FER experimental protocol for design and demographic documentation.
Accounting for Sensor Differences	Untested margin of error for images collected using different thermal sensors.	No mitigation strategy. This is an open research question.	Assessment with optical engineers to determine margin of error across sensors for human thermal FER.

Recommendations

It is daunting to attempt to design a universal, thermal FER benchmark dataset that can account for the myriad of challenges we described. Extensive funding for time, labor, and evaluation would be required. Some challenges are easier to mitigate than others, for example improving the documentation of experimental setup possibly using templates by Gebru et al. (Gebru et al. 2018) and Mitchell et al. (Gebru et al. 2018; Mitchell et al. 2019) versus designing physiological stimuli. But, there may be more feasible short-term solutions that emphasize quality of reviewing the limitations of individual datasets and annotating each with a new labeling system. First, we have observed there are a number of custom datasets as described in Table 2 and are confident that our review missed several proprietary, unpublished, non-English, or classified thermal FER datasets. As a result, there are likely multiple thermal FER databases available all collected with a different set of subjects, experimental setups, and labeling. Offering these in a central online location, would be one step towards inventorying the breadth of data already available worldwide.

Secondly, combining across multiple existing thermal FER datasets and labeling by sensor, domain, posed or spontaneous emotion, resolution, and presence of social context, and stimulus, may be one step towards the aggregation of a larger database. Gathering training data across different datasets is not unusual in thermal FER, as previously noted when Wang et al. (Wang et al. 2014a) combined the NVIE and Equinox datasets to train his DBM model. Both first and second steps would require an effort across researchers to offer up and make available their thermal FER datasets.

Third, despite our review of the thermal FR and FER literature, we struggled to identify any research to evaluate the limits of obfuscating age, gender, ethnicity, and race using thermal imagery. Although some papers affirmed that

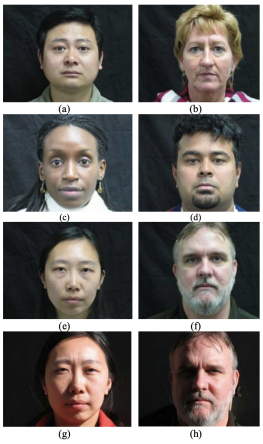

Figure 16: Participants from diverse multimodal dataset collected by the IRIS Lab in 2006 (Chang et al. 2006)

their dataset consisted of diverse demographics (Chang et al. 2006) per Figure 16, none to our knowledge, conducted quantitative tests with human reviewers and inter-rater statistics to test whether or not sensitive demographics could be masked. We believe that in order to assert that thermal imagery can afford any privacy protection and minimize bias, tests must be developed using IRB approval. More broadly, future work should take careful consideration into the scientific questions their research is tackling and the impact it may have in developing or prolonging undesired biases (Friedman and Nissenbaum 1996). Biometrics related research is inherently sensitive and solutions can be valuable to society (Jai 2016). As such researchers should make sure they are familiar with ethical concerns that have occurred in neighboring application areas (Ensign et al. 2018; Chouldechova 2017; Kleinberg, Mullainathan, and Ragha-

van 2016) and remain open to understanding new perspective in which their research may be helpful or detrimental, and could be improved to reduce potential risks (Skirpan and Gorelick 2017; Goldsmith and Burton 2017; Sylvester and Raff 2018).

Conclusion

In this paper, we introduced the advantages of using thermal imagery over RGB for facial FER and provided a survey of thermal FER AI papers, datasets, and selected samples of experimental design protocols. There are several technical benefits of using thermal imagery compared to RGB images for FER, one of which potentially being semi-anonymity. However, there are few labeled, standard thermal affective data sets available for AI training. We have provided a summary of the proposed challenges, with our insights on the consequences, mitigation, and opportunities for each in Table 4.

Acknowledgments

We thank the three anonymous reviewers for AAAI 2020 for their feedback and comments. We also thank Steve Escaravage from Booz Allen Hamilton for his review of this article. This work is supported by grant CRII (IIS–1948399) from the National Science Foundation.

References

Aureli, T.; Grazia, A.; Cardone, D.; and Merla, A. 2015. Behavioral and facial thermal variations in 3-to 4-month-old infants during the still-face paradigm. *Frontiers in psychology* 6:1586.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41(2):423–443.

Bourlai, T.; Pryor, R. R.; Suyama, J.; Reis, S. E.; and Hostler, D. 2012. Use of thermal imagery for estimation of core body temperature during precooling, exertion, and recovery in wildland firefighter protective clothing. *Prehospital Emergency Care*.

Buddharaju, P.; Pavlidis, I. T.; Tsiamyrtzis, P.; and Bazakos, M. 2007. Physiology-based face recognition in the thermal infrared spectrum. *IEEE transactions on pattern analysis and machine intelligence* 29(4):613–626.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91.

Chang, H.; Harishwaran, H.; Yi, M.; Koschan, A.; Abidi, B.; and Abidi, M. 2006. An indoor and outdoor, multimodal, multispectral and multi-illuminant database for face recognition. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 54–54. IEEE.

Chen, C., and Ross, A. 2019. Matching thermal to visible face images using a semantic-guided generative adversarial network. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–8. IEEE.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *FAT ML Workshop*.

Ebisch, S. J.; Aureli, T.; Bafunno, D.; Cardone, D.; Romani, G. L.; and Merla, A. 2012. Mother and child in synchrony: thermal facial imprints of autonomic contagion. *Biological psychology*.

Ekman, P. 1999. Basic emotions. *Handbook of cognition and emotion* 98(45-60):16.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway Feedback Loops in Predictive Policing. In Friedler, S. A., and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 160–171. New York, NY, USA: PMLR.

Equinox. Equinox database at http://www.equinoxsensors.com/products/hid.html.

Fernández-Cuevas, I.; Marins, J. C. B.; Lastras, J. A.; Carmona, P. M. G.; Cano, S. P.; García-Concepción, M. Á.; and Sillero-Quintana, M. 2015. Classification of factors influencing the use of infrared thermography in humans: A review. *Infrared Physics & Technology* 71:28–55.

Friedman, B., and Nissenbaum, H. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14(3):330–347.

Friedrich, G., and Yeshurun, Y. 2002. Seeing people in the dark: Face recognition in infrared images. In *International Workshop on Biologically Motivated Computer Vision*, 348–359. Springer.

Garbey, M.; Sun, N.; Merla, A.; and Pavlidis, I. 2007. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE transactions on Biomedical Engineering* 54(8).

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumeé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Ghiass, R. S.; Arandjelović, O.; Bendada, A.; and Maldague, X. 2014. Infrared face recognition: A comprehensive review of methodologies and databases. *Pattern Recognition* 47(9):2807–2824.

Goldsmith, J., and Burton, E. 2017. Why Teaching Ethics to AI Practitioners Is Important. In *The AAAI-17 workshop on AI, Ethics, and Society*, 110–114.

Goulart, C.; Valadão, C.; Delisle-Rodriguez, D.; Caldeira, E.; and Bastos, T. 2019. Emotion analysis in children through facial emissivity of infrared thermal imaging. *PloS one* 14(3):e0212928.

Greene, D.; Hoffmann, A. L.; and Stark, L. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Grgic, M. Scface - surveillance cameras face database.

Gross, J. J., and Levenson, R. W. 1993. Emotional suppression: physiology, self-report, and expressive behavior. *Journal of personality and social psychology* 64(6):970.

Gunes, H., and Pantic, M. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*.

Hahn, A. C.; Whitehead, R. D.; Albrecht, M.; Lefevre, C. E.; and Perrett, D. I. 2012. Hot or not? thermal reactions to social contact. *Biology letters* 8(5):864–867.

Hamdi, T. 2020. Deepfake detection challenge.

Hammoud, R. I. Otcbvs benchmark dataset collection.

Heo, J.; Kong, S. G.; Abidi, B. R.; and Abidi, M. A. 2004. Fusion of visual and thermal signatures with eyeglass removal for robust face recognition. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 122–122. IEEE.

Hermosilla, G.; Ruiz-del Solar, J.; Verschae, R.; and Correa, M. 2012. A comparative study of thermal face recognition methods in unconstrained environments. *Pattern Recognition* 45(7):2445–2459.

Ioannou, S.; Ebisch, S.; Aureli, T.; Bafunno, D.; Ioannides, H. A.; Cardone, D.; Manini, B.; Romani, G. L.; Gallese, V.; and Merla, A. 2013. The autonomic signature of guilt in children: a thermal infrared imaging study. *PloS one* 8(11):e79440.

Ioannou, S.; Gallese, V.; and Merla, A. 2014. Thermal infrared imaging in psychophysiology: potentialities and limits. *Psychophysiology* 51(10):951–963.

2016. 50 years of biometric research: Accomplishments, challenges, and opportunities.

Jarlier, S.; Grandjean, D.; Delplanque, S.; N'diaye, K.; Cayeux, I.; Velazco, M. I.; Sander, D.; Vuilleumier, P.; and Scherer, K. R. 2011. Thermal analysis of facial muscles contractions. *IEEE transactions on affective computing* 2(1):2–9.

Khan, M. M.; Ingleby, M.; and Ward, R. D. 2006. Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 1(1):91–113.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *FAT ML Workshop*.

Kniaz, V. V.; Knyaz, V. A.; Hladuvka, J.; Kropatsch, W. G.; and Mizginov, V. 2018. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *ECCV*.

Kong, S. G.; Heo, J.; Boughorbel, F.; Zheng, Y.; Abidi, B. R.; Koschan, A.; Yi, M.; and Abidi, M. A. 2007. Multiscale fusion of visible and thermal ir images for illumination-invariant face recognition. *International Journal of Computer Vision* 71(2):215–233.

Kopaczka, M.; Kolk, R.; and Merhof, D. 2018. A fully annotated thermal face database and its application for thermal facial expression recognition. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 1–6. IEEE.

Kosonogov, V.; De Zorzi, L.; Honore, J.; Martínez-Velázquez, E. S.; Nandrino, J.-L.; Martinez-Selva, J. M.; and Sequeira, H. 2017. Facial thermal variations: A new marker of emotional arousal. *PloS one* 12(9):e0183592.

Kumar, A. Iit delhi near ir face database version 2.0.

Levenson, R. W. 1988. Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. *Social psychophysiology: Theory and clinical applications*.

Li, S., and Deng, W. 2018. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*.

Lohr, S. 2018. Facial recognition is accurate, if you're a white guy.

Lopez, M. B.; del Blanco, C. R.; and Garcia, N. 2017. Detecting exercise-induced fatigue using thermal imaging and deep learning. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. IEEE.

Mallat, K., and Dugelay, J.-L. 2018. A benchmark database of visible and thermal paired face images across multiple variations. In *2018 BIOSIG*, 1–5. IEEE.

Mallat, K.; Damer, N.; Boutros, F.; Kuijper, A.; and Dugelay, J.-L. 2019. Cross-spectrum thermal to visible face recognition based on cascaded image synthesis. In *2019 International Conference on Biometrics (ICB)*, 1–8. IEEE.

Manini, B.; Cardone, D.; Ebisch, S.; Bafunno, D.; Aureli, T.; and Merla, A. 2013. Mom feels what her child feels: thermal signatures of vicarious autonomic response while watching children in a stressful situation. *Frontiers in human neuroscience* 7:299.

Martinez-Martin, N. 2019. What are important ethical implications of using facial recognition technology in health care? *AMA journal of ethics* 21(2):E180.

Matsakis, L. 2020. Amazon won't let police use its facial-recognition tech for one year.

McDuff, D.; Girard, J. M.; and El Kaliouby, R. 2017. Large-scale observational evidence of cross-cultural differences in facial behavior. *Journal of Nonverbal Behavior* 41(1):1–19.

Merla, A.; Di Donato, L.; Rossini, P.; and Romani, G. 2004. Emotion detection through functional infrared imaging: preliminary results. *Biomedizinische Technick* 48(2):284–286.

Merla, A. 2014. Revealing psychophysiology and emotions through thermal infrared imaging. In *PhyCS*, 368–377.

Mesquita, B., and Boiger, M. 2014. Emotions in context: A sociodynamic model of emotions. *Emotion Review* 6(4):298–302.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Murgia, M. 2019. Microsoft quietly deletes largest public face recognition data set.

Nguyen, D. T., and Park, K. R. 2016. Body-based gender recognition using images from visible and thermal cameras. *Sensors* 16(2):156.

Nguyen, H.; Kotani, K.; Chen, F.; and Le, B. 2013. A thermal facial emotion database and its analysis. In *Pacific-Rim Symposium on Image and Video Technology*. Springer.

Nhan, B. R., and Chau, T. 2009. Classifying affective states using thermal infrared imaging of the human face. *IEEE Transactions on Biomedical Engineering* 57(4):979–987.

Panetta, K.; Wan, Q.; Agaian, S.; Rajeev, S.; Kamath, S.; Rajendran, R.; Rao, S.; Kaszowska, A.; Taylor, H.; Samani, A.; et al. 2018. A comprehensive database for benchmarking imaging systems. *IEEE transactions on pattern analysis and machine intelligence*.

Pavlidis, I., and Symosek, P. 2000. The imaging issue in an automatic face/disguise detection system. In *Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No. PR00640)*, 15–24. IEEE.

Pavlidis, I.; Dowdall, J.; Sun, N.; Puri, C.; Fei, J.; and Garbey, M. 2007. Interacting with human physiology. *Computer Vision and Image Understanding* 108(1-2):150–170.

Pavlidis, I.; Tsiamyrtzis, P.; Shastri, D.; Wesley, A.; Zhou, Y.; Lindner, P.; Buddharaju, P.; Joseph, R.; Mandapati, A.; Dunkin, B.; et al. 2012. Fast by nature-how stress patterns define human experience and performance in dexterous tasks. *Scientific Reports* 2:305.

Pittaluga, F.; Zivkovic, A.; and Koppal, S. J. 2016. Sensor-level privacy for thermal cameras. In *2016 IEEE International Conference on Computational Photography (ICCP)*, 1–12. IEEE.

Puri, C.; Olson, L.; Pavlidis, I.; Levine, J.; and Starren, J. 2005. Stresscam: non-contact measurement of users' emotional states through thermal imaging. In *CHI'05 extended abstracts on Human factors in computing systems*.

Raff, E. 2019. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, 5485–5495.

Reid, D. A.; Samangooei, S.; Chen, C.; Nixon, M. S.; and Ross, A. 2013. Soft biometrics for surveillance: an overview. In *Handbook of statistics*, volume 31. Elsevier. 327–352.

Salazar-López, E.; Domínguez, E.; Ramos, V. J.; De la Fuente, J.; Meins, A.; Iborra, O.; Gálvez, G.; Rodríguez-Artacho, M.; and Gómez-Milán, E. 2015. The mental and subjective skin: Emotion, empathy, feelings and thermography. *Consciousness and cognition*.

Seo, J., and Chung, I.-J. 2019. Face liveness detection using thermal face-cnn with external knowledge. *Symmetry* 11(3):360.

Shoja Ghiass, R. 2014. Face recognition using infrared vision.

Simón, M. O.; Corneanu, C.; Nasrollahi, K.; Nikisins, O.; Escalera, S.; Sun, Y.; Li, H.; Sun, Z.; Moeslund, T. B.; and Greitans, M. 2016. Improved rgb-dt based face recognition. *Iet Biometrics* 5(4):297–303.

Singer, N., and Metz, C. 2019. Many facial-recognition systems are biased, says u.s. study.

Skirpan, M., and Gorelick, M. 2017. The Authority of "Fair" in Machine Learning. In *FAT ML Workshop*.

Sonkusare, S.; Ahmedt-Aristizabal, D.; Aburn, M. J.; Nguyen, V. T.; Pang, T.; Frydman, S.; Denman, S.; Fookes, C.; Breakspear, M.; and Guo, C. C. 2019. Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. *Scientific reports* 9(1):1–11.

Sylvester, J., and Raff, E. 2018. What About Applied Fairness? In *Machine Learning: The Debates (ML-D) organized as part of the Federated AI Meeting (FAIM 2018)*.

Tian, Y.-L.; Kanade, T.; and Cohn, J. F. 2005. Facial expression analysis. In *Handbook of face recognition*. Springer. 247–275.

Ting, D. S. W.; Carin, L.; Dzau, V.; and Wong, T. Y. 2020. Digital technology and covid-19. *Nature medicine* 26(4):459–461.

Trujillo, L.; Olague, G.; Hammoud, R.; and Hernandez, B. 2005. Automatic feature localization in thermal images for facial expression recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE.

UND. University of notre dame und-collection c.

Van Natta, M.; Chen, P.; Herbek, S.; Jain, R.; Kastelic, N.; Katz, E.; Struble, M.; Vanam, V.; and Vattikonda, N. 2020. rise and regulation of thermal facial recognition technology during the covid-19 pandemic.

Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; and Wang, X. 2010. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia* 12(7):682–691.

Wang, S.; He, M.; Gao, Z.; He, S.; and Ji, Q. 2014a. Emotion recognition from thermal infrared images using deep boltzmann machine. *Frontiers of Computer Science* 8(4):609–618.

Wang, S.; He, S.; Wu, Y.; He, M.; and Ji, Q. 2014b. Fusion of visible and thermal images for facial expression recognition. *Frontiers of Computer Science* 8(2):232–242.

Wilder, J.; Phillips, P. J.; Jiang, C.; and Wiener, S. 1996. Comparison of visible and infra-red imagery for face recognition. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 182–187. IEEE.

Yan, W.-J.; Li, X.; Wang, S.-J.; Zhao, G.; Liu, Y.-J.; Chen, Y.-H.; and Fu, X. 2014. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one* 9(1):e86041.

Yoshitomi, Y.; Kim, S.-I.; Kawano, T.; and Kilazoe, T. 2000. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499)*. IEEE.

Zhang, B.; Zhang, L.; Zhang, D.; and Shen, L. 2010. Directional binary code with application to polyu near-infrared face database. *Pattern Recognition Letters* 31(14):2337–2344.

Zhang, T.; Wiliem, A.; Yang, S.; and Lovell, B. 2018. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 ICB*. IEEE.

Zhu, Z.; Tsiamyrtzis, P.; and Pavlidis, I. 2007. Forehead thermal signature extraction in lie detection. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 243–246. IEEE.