

MSX-FTID: A Novel Multi-spectral Dynamic Facial Thermal Image Database for Emotion Recognition

Sakshi Gupta

521ee1010@nitrrkl.ac.in

National Institute of Technology

Anwesha Sengupta

National Institute of Technology

Research Article

Keywords: Emotion recognition, Thermal images, MSX, Database, Expressions

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8800893/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

MSX-FTID: A Novel Multi-spectral Dynamic Facial Thermal Image Database for Emotion Recognition

Sakshi Gupta* and Anwesha Sengupta

Department of Electrical Engineering, National Institute of Technology, Rourkela,
769008, Odisha, India.

*Corresponding author(s). E-mail(s): iamsakshi.usha@gmail.com;
Contributing authors: senguptaan@nitrkl.ac.in;

Abstract

Analysis of facial expressions provide valuable insight into emotional states of subjects. Understanding human emotions is fundamental for applications such as affective computing, human-computer interaction, and healthcare. Conventional methods for facial expression analysis rely primarily on visible-light imaging, which can be vulnerable to misinterpretation. Recent advancements in thermal imaging technology have opened new avenues for non-invasive emotion recognition through the analysis of facial temperature variations. The present study proposes a spontaneous multi-spectral dynamic facial thermal image database (MSX-FTID) and validates the same using a deep learning-based approach for six basic standard emotions. The validation results demonstrate the efficacy of the proposed database in recognizing human emotions accurately, showcasing its potential for real-world applications in affective computing, healthcare, and human-computer interaction.

Keywords: Emotion recognition, Thermal images, MSX, Database, Expressions

1 Introduction

Emotion Recognition (ER) finds wide application in fields such as human-machine interaction (HMI), affective computing, mental health monitoring, personalized learning systems, marketing analytics, and security systems [1]. ER enhances empathetic and efficient human-centered interactions by enabling machines to perceive and interpret human emotional states. Contemporary ER systems rely on multimodal data sources, including physiological signals (e.g., heart rate variability and galvanic skin response), neurophysiological signals such as electroencephalography (EEG), speech prosody, and behavioral cues like eye tracking, facial expressions, body posture,

and gestures [2][3]. Integrating multiple modalities has been shown to improve robustness and accuracy, extending ER applications to areas such as emotion-aware virtual assistants, adaptive user interfaces, and real-time behavioral health diagnostics [4][5].

Facial imaging remains one of the most effective and non-intrusive approaches for emotion recognition [6]. The majority of recent studies have focused on visible-spectrum images or videos, leveraging the widespread availability of CCD (Charge-Coupled Device) and CMOS (Complementary Metal-Oxide-Semiconductor) cameras and the success of deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for analyzing

static and dynamic facial expressions [7][8]. These methods often rely on facial Action Units (AUs) derived from the Facial Action Coding System (FACS) to classify emotions. However, visible-spectrum imaging suffers from inherent limitations, including sensitivity to illumination variations, shadows, and low-light conditions, which can significantly degrade performance [9][10][11]. Furthermore, visible images primarily capture surface-level facial features, limiting their ability to detect subtle or involuntary emotional cues.

Infrared thermal imaging has emerged as a promising alternative to address such limitations by capturing facial heat radiation associated with physiological and autonomic nervous system (ANS) responses [11]. Unlike visible imaging, thermal imaging is largely invariant to lighting conditions and is less susceptible to voluntary control, making it particularly suitable for emotion recognition in unconstrained environments [12]. Operating in the long-wavelength infrared (LWIR) range (7.5–14 μm), thermal cameras measure temperature variations related to blood flow and refer to emotional arousal [13]. Facial regions such as the forehead, nose, cheeks, and periorbital areas exhibit temperature changes corresponding to emotional states like stress, fear, and excitement [14]. However, thermal images often lack fine-grained structural details, which can hinder precise facial feature localization [15].

Multi-Spectral Dynamic (MSX) imaging, which offers a combination of thermal and visible-spectrum information to enhance spatial detail while retaining physiological cues, has hence been proposed to circumvent this problem. MSX technology improves the reliability and interpretability of emotion recognition systems [16] by combining the robustness of thermal imaging with the high-resolution structural information of visible images. Such multimodal integration enables more accurate emotion analysis across diverse environmental conditions and supports applications in affective computing, healthcare, surveillance, and security.

Existing emotion recognition research is mostly reliant on posed expression databases, wherein participants are instructed to display specific emotions. In contrast, spontaneous expressions elicited by specific stimuli are more subtle and exhibit natural temporal dynamics, including

micro-expressions and involuntary facial movements [17]. To the best of the knowledge of the authors, there is limited availability of spontaneous expression datasets incorporating thermal images. Development of spontaneous expression datasets is essential to facilitate natural emotion recognition, thereby enabling further practical, real-world applications.

This paper attempts to advance research in facial expression recognition using thermal imaging by proposing the Multi-Spectral Dynamic Facial Thermal Image Database (MSX-FTID). The proposed database is designed to support the evaluation of emotion recognition systems under realistic conditions and addresses key challenges associated with thermal data acquisition and annotation. The dataset comprises six fundamental emotions, namely happiness, sadness, surprise, disgust, fear, and anger, elicited through emotion-inducing video stimuli to capture spontaneous facial responses. All samples are manually segmented and annotated with relevant metadata, including expression intensity, gender, and facial occlusions. Emotion recognition is performed using deep learning-based facial thermal analysis.

The main contributions of this work are summarized as follows:

1. This work demonstrates the effectiveness of MSX-based facial thermal imaging for recognizing six basic human emotions.
2. A novel spontaneous thermal facial emotion dataset is introduced to address the lack of thermal emotion databases for the Indian ethnicity.
3. The dataset is comprehensively validated by benchmarking it across multiple standard pre-trained deep learning architectures, demonstrating its robustness and suitability for emotion recognition research.

The rest of the paper is organized as follows: Section II briefly reviews the existing emotion databases. Sections III and IV show the experiment design and the database content, respectively. Section V describes the evaluation of the database. Sections VI and VII describe the limitations and conclusions of this work respectively.

2 Brief Review of Existing Databases

Research on facial expression databases has evolved significantly since the late 1990s, with many well-known datasets such as JAFFE [18] and RAVDEES [19] primarily containing posed expressions. However, spontaneous expressions provide more reliable cues for emotion recognition due to natural variations in intensity and temporal dynamics. Over time, several visible-spectrum databases focusing on spontaneous expressions have been developed, including UA-UIUC [20], MMI [21], UT-Dallas [22], AvID [23], the Geneva Airport Lost Luggage Study [24], Belfast [25], and AAI [26]. However, an analysis of the databases highlights that collecting high-quality spontaneous emotional data across diverse cultural contexts remains challenging, particularly for advancing research in thermal-based emotion recognition [9].

In contrast to the large number of visible-spectrum datasets, only a limited number of thermal facial expression databases exist, as summarized in Table 1. Some widely referenced datasets, such as the NIST Equinox database, are no longer publicly available, while others, including the IRIS Thermal/Visible Face Dataset [27], primarily focus on posed expressions and are more suited for face recognition than spontaneous emotion analysis. Very few datasets incorporate both visible and thermal modalities for studying spontaneous emotional responses. To address unimodal limitations, MSX-based imaging has been explored for improved emotion recognition [16], detailed in Table 2. However, existing MSX datasets consider only five emotional categories and rely on less effective elicitation methods [35][36], highlighting the need for a standardized dataset covering all six universal emotions. These gaps motivated the development of the proposed MSX thermal facial emotion dataset.

3 Experiment Design

3.1 Choice of Participants

Seventeen participants, comprising 10 males and 7 females, aged between 20 and 30 years (mean age 23 ± 2 years), were chosen as volunteers from graduate and post-graduate students of NIT Rourkela.

The inclusion and exclusion criteria for the subjects have been shown in Table 3. The participants were required to possess normal auditory acuity. The participants were required to fill out the General Health Questionnaire (GHQ [37]) and the State-Trait Anxiety Inventory (STAI [38]) prior to the commencement of the experiment.

The participants were informed about the time of the experiment a day in advance. The participants had the right to withdraw from the experiment at any point of time.

3.2 Experimental setup

The experiment was conducted in a $4m \times 5m$ enclosed room with a single entry point to ensure isolation and minimize external disturbances. Participants were not informed of the study’s objective beforehand to promote spontaneous emotional responses and reduce the likelihood of biased or artificial behavior. Facial images were recorded using a camera concealed within a sealed enclosure, minimizing participants’ awareness of being observed. This arrangement helped eliminate the observer effect [20], which can lead to exaggerated or masked expressions when subjects are conscious of recording equipment. The overall experimental setup is illustrated in Fig. 1.

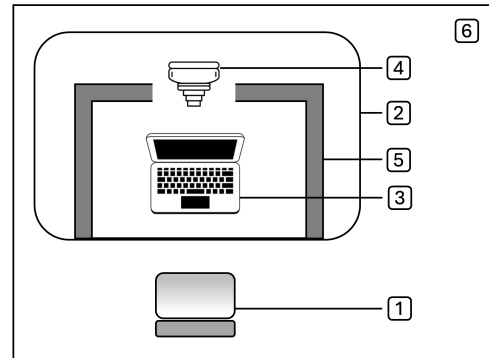


Fig. 1 The Experimental Set-up: (1) chair for subject, (2) table, (3) laptop, (4) infrared camera, (5) box to conceal camera, (6) experiment room

MSX thermal images were captured using a FLIR TG165-X infrared camera with an IR resolution of 80×60 pixels. The camera was positioned approximately 2 meters above the ground at a distance of 1–1.5 meters from the participant. To ensure accurate temperature measurement, the camera was calibrated prior to each session.

Table 1 Existing facial thermal databases

Database (Year)	Subjects	Data description	Waveband	Elicitation	Elicitation method	Expressions
NIST/Equinox	600	1919 thermal images	8-13 μ m	Posed	N/A	Smile, frowning, and surprise
IRIS [27] (2004)	30	4228 pairs of visible and thermal images	7-14 μ m	Posed	N/A	Surprise, laughing, and anger
NVIE [28] (2010)	215	Visible and thermal images	8-14 μ m	Posed and Spontaneous	Emotional videos	Happy, angry, sad, disgust, fearful, neutral and surprise
UCHT [29] (2012)	102	Thermal images	7.5-13.5 μ m	Posed	N/A	Happy, sad, and angry
KTFE [30] (2013)	26	2539 pairs of visible and thermal images	8-14 μ m	Spontaneous	Emotional videos	Happy, angry, sad, disgust, fearful, neutral and surprise
RGB-DT [31] (2014)	51	Synchronized RGB, depth and thermal face images	7.5-14 μ m	Posed	N/A	Happy, angry, sad, surprise, and neutral
RWTH Aachen University [32] (2017)	82	Thermal videos	7-14 μ m	Posed	N/A	Neutral, Happy, sad, disgust, fearful, angry, surprise, and contempt
VIS-TH [33] (2018)	50	4200 visible and thermal images	7.5-13.5 μ m	Posed	N/A	Happy, angry, sad, surprise, and neutral
Tufts [34] (2020)	113	Photographic images, a computerized facial sketch, and thermal images	7.5 – 14 μ m	Posed	N/A	Smile, neutral, and shocked

Note: N/A tends not to be applicable to the elicitation method, as those databases have posed expressions only.

Table 2 Existing MSX facial thermal databases

Database (Year)	Subjects	Data description	Waveband	Elicitation	Elicitation method	Expressions
VIRI [16] (2020)	110	1698 visible, IR and MSX thermal images	8 – 14 μ m	Spontaneous	Emotion elicitation questions	Happy, angry, neutral, sad and surprise

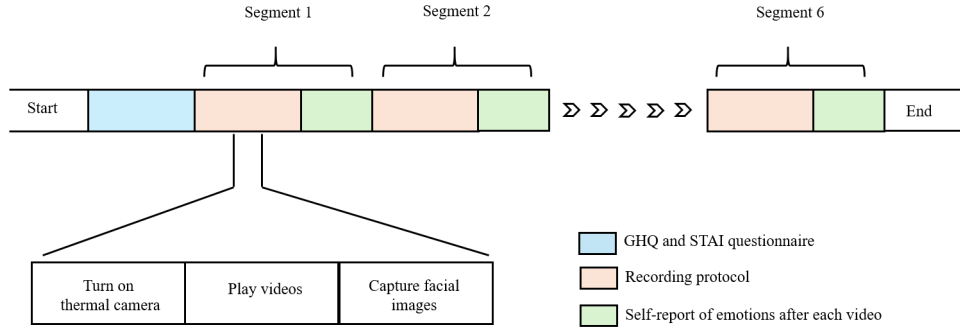
3.2.1 Recording procedure

The data collection process was conducted with only one participant and one experimenter present in the room at a time to ensure optimal emotional engagement and minimize distractions. Participants were seated comfortably in front of a laptop screen without physical restrictions, as shown in Fig. 1. Emotion-eliciting video clips were displayed on a 26-inch LCD screen in six separate segments, each corresponding to a specific emotion. Facial expressions were captured as a sequence of images at regular intervals to record the gradual transition of emotions during the elicitation task.

A 2-minute interval was maintained between successive segments to reduce emotional overlap. During this period, participants identified the emotions they experienced and rated their intensity on a scale of 1 to 5. These self-reports were used by the experimenter to label the captured images. The overall recording procedure is illustrated in Fig. 2. After the experiment, participants were debriefed about the study objectives and provided written consent for the use of their facial images. Ethical approval for the study was granted by the Institute Ethical Committee (IEC) of NIT Rourkela.

Table 3 Inclusion and exclusion criteria of subjects for the experiment

Inclusion criteria	Exclusion Criteria
1. Age range 20 to 30 years	1. Previously participated
2. Normal/corrected vision and/or hearing ability	2. Sleep disorder
3. Adequately rested for the past 2 hours	3. History of substance and/or psychotropic drug abuse
4. Adequate sleep of 6 to 8 hours the night before	4. Having any chronic disease or long-term medications
5. No history of neurological or psychiatric disorders	

**Fig. 2** Recording procedure

3.2.2 Choice of Stimuli

Eliciting emotions through video clips is a well-established approach in affective research, as dynamic audiovisual stimuli effectively evoke spontaneous emotional responses. Carvalho et al. [36] demonstrated that videos are more effective than static images or audio alone in inducing genuine emotional states. In this study, video clips were carefully selected to align with the Indian demographic, ensuring cultural relevance. Drawing from established resources such as the Emotional Movie Database (EMDB) [36] and prior studies on film-based emotion elicitation [39], the clips were categorized into six emotional classes: happiness, surprise, fear, sadness, disgust, and anger [17].

The duration of the clips ranged from 10 seconds to 3 minutes, following recommended guidelines for effective emotion induction while minimizing participant fatigue [40]. Pre-validation ensured the suitability of the clips for eliciting the intended emotional responses, enhancing the study's ecological validity and cultural appropriateness. The distribution of video categories is summarized in Table 4, providing a structured

framework for analyzing affective responses across participants.

Table 4 Emotion elicitation videos of each category

Emotions	Number of videos
Happy	3
Surprise	2
Fear	3
Sad	5
Disgust	4
Angry	2

4 Database Content

The proposed Multi-Spectral Dynamic Facial Thermal Image Database (MSX-FTID) contains 3562 MSX facial thermal images of 17 subjects depicting six spontaneous emotions. The recording time for each participant was around 30 minutes. The dataset details are provided in Table 5. It also shows the image counts of each emotion class in the dataset.

The proposed database contains subtle to full-blown elicitation of different emotions. A few

Table 5 Summary of Database Content

Number of facial images	3562 MSX thermal images
Number of participants	17 (10 male and 7 female)
Emotions elicited	Happy (256 images), Surprise(304 images), Fear (1045 images), Sad (973 images), Disgust (579 images), Angry (405 images)
Self-report of emotion	Yes
Emotion rating scale	1-5 (1: No emotion, 5: Maximum emotion intensity)

examples of full-blown images from the database are provided in Fig. 3. It shows the six different emotions of the subjects. Different expressions of the same subject are shown in Fig. 4.

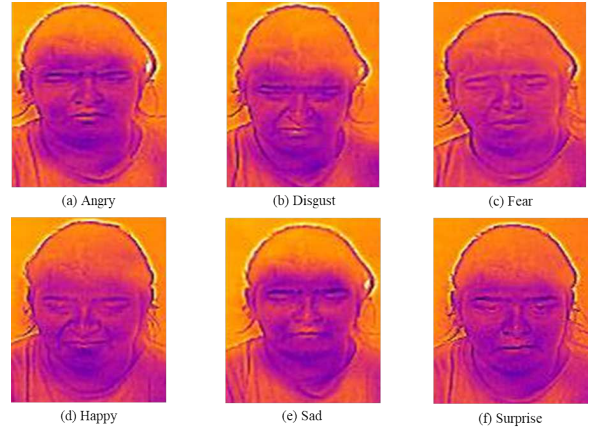
For each emotion class, the image sequences for each subject begin with a neutral facial expression and progress to either the peak intensity of the emotion or its offset. This structured progression enables researchers to effectively analyze and model the dynamic transition of spontaneous facial expressions from a neutral baseline to the emotional apex and resolution. The database is available on request via email to the authors.

5 Results & Analysis

The main aim of this section is to present an elementary assessment of the usability of the spontaneous thermal database for expression recognition and emotion inference, and to provide reference evaluation results for researchers using the database.

5.1 Performance Comparison

To validate the robustness of the proposed MSX Facial Thermal Emotion Dataset (MSX-FTID), multiple convolutional and transformer-based architectures were benchmarked on the

**Fig. 3** Example images of six expressions (Angry, Disgust, Fear, Happy, Sad, Surprise)**Fig. 4** Example images of a subject showing the six expressions

KTFE [30] dataset and compared with the proposed dataset to assess their generalizability and recognition difficulty.

Each model was trained and tested using the same experimental settings, including identical image preprocessing, batch size, learning rate, and data split to ensure a fair comparison. The models were evaluated using categorical cross-entropy loss and accuracy as the primary metrics. Table 6 summarizes the accuracy results obtained for the six basic emotion classes using different pretrained and baseline architectures.

While all models were evaluated under identical conditions, MobileNetV2 achieved the highest validation accuracy of 81.6%, followed by ResNet50 (75%), VGG-16 (72.3%), and Vision Transformer (52.86%). These models demonstrate better feature abstraction and robustness in capturing subtle thermal variations across facial regions.

This visual comparison supports the findings in Table 6 and validates the efficiency, stability, and benchmarking potential of the proposed dataset across diverse model architectures.

MobileNetV2 Model recognizes emotions through facial expressions with an accuracy of 81.60 %. The accuracy and loss graphs for the emotions classified using our proposed dataset are shown in Fig. 5 and Fig. 6, respectively. The X-axis of the graphs reflects the number of epochs conducted, while the Y-axis represents the model's accuracy and the model's loss, respectively.

When evaluated on the dataset, the model's accuracy grew for 50 epochs before converging. However, there are some peaks and valleys in the midst. After 50 epochs, the loss also begins to converge. Examining the accuracy and loss curves on the proposed datasets demonstrates that the model developed in this work has strong generalization and resilience.

This outcome indicates the effectiveness of our approach in recognizing and detecting the facial emotion of interest within the proposed dataset. These results pave the way for further advancements in emotion recognition and detection systems.

6 Limitation and Challenges

During the data collection process for facial expression recognition using thermal imaging, several challenges were encountered. Unlike visible-spectrum imaging, thermal imaging captures

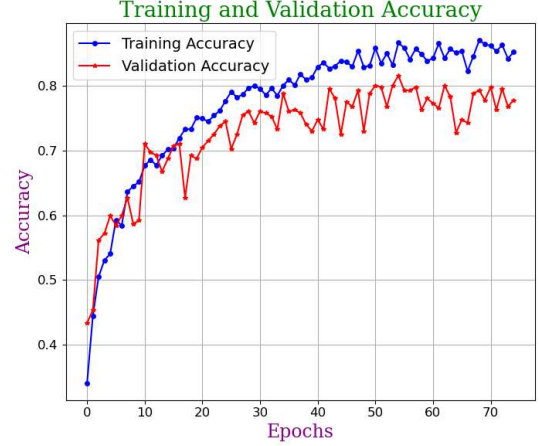


Fig. 5 Training Vs. Validation Accuracy

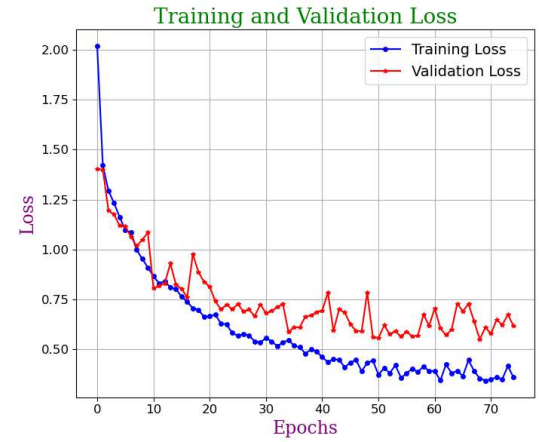


Fig. 6 Training Vs. Validation Loss

physiological signals, including facial temperature variations influenced by both muscular activity and autonomic responses. This complexity introduces variability in consistently capturing emotional expressions. Additional challenges arose in visualizing thermal data due to infrared artefacts and palette inconsistencies. To address this, standardized colour mapping schemes were applied during dataset compilation, ensuring visual consistency and improved interpretability.

Despite these limitations, the dataset offers valuable insights into how facial temperature distributions correlate with emotional states. It advances the field of emotion recognition by integrating thermal imaging with facial analysis, paving the way for applications in affective

Table 6 Accuracy comparison of various standard models on MSX-FTID Dataset

Sr.No.	Model	Architecture type	Accuracy (KTFE)	Accuracy (Proposed)
1	Vision Transformer (ViT) [41]	Transformer-based	47.00 %	52.86 %
2	VGG-16 [42]	Deep CNN	29.30 %	72.30 %
3	ResNet50 [43]	Residual CNN (Deeper)	55.47 %	75.00 %
4	MobileNetV2 [44]	Lightweight Depthwise CNN	63.50 %	81.60 %

computing, healthcare diagnostics, psychological assessment, and human-computer interaction.

7 Conclusion and Future Scope

This study highlights the potential of thermal imaging as a robust alternative to conventional visible light imaging for emotion recognition. By leveraging a novel multi-spectral dynamic facial thermal image dataset and employing a deep learning-based approach, the present work demonstrates superior accuracy in classifying human emotions. The results validate the efficacy of combining visual and thermal imaging modalities, showcasing their applicability in real-world scenarios such as affective computing, healthcare, and human-computer interaction. Facial images and expressions vary across cultures; thus, this method may be limited to individuals of a particular ethnicity and real-time human monitoring. Future work will focus on further enhancing the dataset and improving the generalizability of the proposed model to diverse populations and environments, paving the way for more reliable and versatile emotion recognition systems.

Acknowledgment

The authors want to acknowledge the help extended by Rajat Verma, Jainam Doshi, Gaurav Mahajan, and all the participants of NIT Rourkela.

Declarations

- **Funding:** Not applicable
- **Conflict of interest:** The authors declare that they have no conflict of interest.

- **Ethics approval:** Ethical approval was granted by the Institute Ethical Committee (IEC) of NIT Rourkela.
- **Consent to participate:** Participants provided written consent to use their data for research purposes.
- **Consent for publication:** Participants provided written consent to publish their facial images.
- **Availability of data and materials:** The dataset is available upon request via email at 521ee1010@nitrkl.ac.in and senguptaan@nitrkl.ac.in.
- **Authors' contributions:** S.G. was responsible for experimentation, writing, and editing of the article. A.S. helped in guiding the necessary changes and ideas required for fine-tuning results, along with the final proofreading of the article.

References

- [1] Ekman, P. What scientists who study emotion agree about. *Perspectives on psychological science* **11**, 31–34 (2016).
- [2] Dziedzickis, A., Kaklauskas, A. & Bucinskas, V. Human emotion recognition: Review of sensors and methods. *Sensors* **20**, 592 (2020).
- [3] Fasel, B. & Luetttin, J. Automatic facial expression analysis: a survey. *Pattern recognition* **36**, 259–275 (2003).
- [4] Li, X. & Lin, R. Speech emotion recognition for power customer service (2021).
- [5] Simcock, G. *et al.* Associations between facial emotion recognition and mental health in early adolescence. *International journal of environmental research and public health* **17**, 330 (2020).

- [6] Ekman, P. Emotions revealed: recognizing faces and feelings to improve communication and emotional life. new york. *NY: Times books* (2003).
- [7] Li, S. Z., Jain, A. K., Tian, Y.-L., Kanade, T. & Cohn, J. F. Facial expression analysis. *Handbook of face recognition* 247–275 (2005).
- [8] Fan, Y., Lu, X., Li, D. & Liu, Y. Video-based emotion recognition using cnn-rnn and c3d hybrid networks (2016).
- [9] Ko, B. C. A brief review of facial emotion recognition based on visual information. *sensors* **18**, 401 (2018).
- [10] Canal, F. Z. *et al.* A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences* **582**, 593–617 (2022).
- [11] Khan, M. M., Ward, R. D. & Ingleby, M. Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. *ACM Transactions on Applied Perception (TAP)* **6**, 1–22 (2009).
- [12] Krzywicki, A. T., He, G. & O’Kane, B. L. Analysis of facial thermal variations in response to emotion: eliciting film clips (2009).
- [13] Rooj, S., Routray, A. & Mandal, M. K. Feature based analysis of thermal images for emotion recognition. *Engineering Applications of Artificial Intelligence* **120**, 105809 (2023).
- [14] Bhowmik, M. K., Bhattacharjee, D., Nasipuri, M., Basu, D. K. & Kundu, M. Optimum fusion of visual and thermal face images for recognition (2010).
- [15] James, A. P. & Dasarathy, B. V. Medical image fusion: A survey of the state of the art. *Information Fusion* **19**, 4–19 (2014). URL <https://www.sciencedirect.com/science/article/pii/S1566253513001450>. Special Issue on Information Fusion in Medical Image Computing and Systems.
- [16] Siddiqui, M. F. H., Dhakal, P., Yang, X. & Javaid, A. Y. A survey on databases for multimodal emotion recognition and an introduction to the viri (visible and infrared image) database. *Multimodal Technologies and Interaction* **6**, 47 (2022).
- [17] Happy, S., Patnaik, P., Routray, A. & Guha, R. The indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing* **8**, 131–142 (2015).
- [18] Shih, F. Y., Chuang, C.-F. & Wang, P. S. Performance comparisons of facial expression recognition in jaffe database. *International Journal of Pattern Recognition and Artificial Intelligence* **22**, 445–459 (2008).
- [19] Livingstone, S. R. & Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one* **13**, e0196391 (2018).
- [20] Sebe, N. *et al.* Authentic facial expression analysis. *Image and vision computing* **25**, 1856–1863 (2007).
- [21] Pantic, M. & Bartlett, M. S. Machine analysis of facial expressions (2007).
- [22] O’Toole, A. J. *et al.* A video database of moving faces and people. *IEEE Transactions on pattern analysis and machine intelligence* **27**, 812–816 (2005).
- [23] Gajšek, R. *et al.* Multi-modal emotional database: Avid. *Informatica* **33** (2009).
- [24] Scherer, K. R. & Ceschi, G. Lost luggage: a field study of emotion–antecedent appraisal. *Motivation and emotion* **21**, 211–235 (1997).
- [25] Sneddon, I., McRorie, M., McKeown, G. & Hanratty, J. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing* **3**, 32–41 (2011).

- [26] Roisman, G. I., Tsai, J. L. & Chiang, K.-H. S. The emotional integration of childhood experience: physiological, facial expressive, and self-reported emotional response during the adult attachment interview. *Developmental psychology* **40**, 776 (2004).
- [27] Hammoud, D. R. I. Dataset 02: Iris thermal/visible face database. *OTCBVS Benchmark Dataset Collection* (2004). URL <http://vcip-okstate.org/pbvs/bench/Data/02/download.html>.
- [28] Wang, S. *et al.* A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia* **12**, 682–691 (2010).
- [29] Hermosilla, G., Ruiz-del Solar, J., Verschae, R. & Correa, M. A comparative study of thermal face recognition methods in unconstrained environments. *Pattern Recognition* **45**, 2445–2459 (2012).
- [30] Nguyen, H., Kotani, K., Chen, F. & Le, B. A thermal facial emotion database and its analysis (2014).
- [31] Nikisins, O., Nasrollahi, K., Greitans, M. & Moeslund, T. B. Rgb-dt based face recognition (2014).
- [32] Kopaczka, M., Nestler, J. & Merhof, D. Face detection in thermal infrared images: A comparison of algorithm-and machine-learning-based approaches (2017).
- [33] Mallat, K. & Dugelay, J.-L. A benchmark database of visible and thermal paired face images across multiple variations (2018).
- [34] Panetta, K. *et al.* A comprehensive database for benchmarking imaging systems. *IEEE transactions on pattern analysis and machine intelligence* **42**, 509–520 (2018).
- [35] Dhaka, S. & Kashyap, N. Explicit emotion regulation: Comparing emotion inducing stimuli. *Psychological Thought* **10** (2017).
- [36] Carvalho, S., Leite, J., Galdo-Álvarez, S. & Gonçalves, O. F. The emotional movie database (emdb): A self-report and psychophysiological study. *Applied psychophysiology and biofeedback* **37**, 279–294 (2012).
- [37] Goldberg, D. P. & Hillier, V. F. A scaled version of the general health questionnaire. *Psychological medicine* **9**, 139–145 (1979).
- [38] Cattell, R. Charles d. spielberger. *Anxiety: Current trends in theory and research* 23 (2013).
- [39] Schaefer, A., Nils, F., Sanchez, X. & Philippot, P. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion* **24**, 1153–1172 (2010).
- [40] Ray, R. & Gross, J. Emotion elicitation using films in handbook of emotion elicitation and assessment (eds coan, a. & allen, jj b.) 9–28 (2007).
- [41] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
- [42] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [43] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition (2016).
- [44] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks (2018).