RESEARCH PAPER

Deep Multimodal Fusion of Thermal and Visible Images for Emotion Interpretation

Dr. Aparna A. Joshi[1]* , Dr. Rita B Patil [2], Dr. Rasika Kulkarni [3], Dr. Almas M. N. Siddiqui [4],

1. Assistant Professor, Pratibha College of Commerce and Computer Studies, Chinchwad, Pune, M.S., India. aajoshi2@gmail.com
2. Assistant Professor, Dr. G. Y. Pathrikar college of CS and IT, MGM University, Chh. Sambhajinagar, M.S., India. ritapatil@gmail.com

3.Assistant Professor, Fergusson College, Pune, M.S., India. rasika.kulkarni @fergusson.com.

4. Assistant Professor, Vivekanand Institute of Advanced Studies in Management Science and Communications, Chh. Sambhajinagar, M.S., India. siddiqui.almas29@gmail.com.

Corresponding Author: Dr. Aparna A. Joshi[1]* , Assistant Professor, Pratibha College of Commerce and Computer Studies, Chinchwad, Pune, M.S., India. E mail: Aajoshi2@gmail.com

Abstract

Emotion recognition plays a crucial role in affective computing and human–computer interaction, enabling intelligent systems to understand and respond to human emotional states. Conventional emotion recognition approaches primarily rely on visible-light facial images; however, their performance is significantly affected by variations in illumination, occlusion, and low-light conditions. Thermal imaging offers a complementary modality by capturing physiological changes such as blood flow and temperature variations associated with emotional responses, and it remains robust under challenging environmental conditions. This study explores emotion interpretation using fused thermal and visible facial images through deep learning techniques. A multimodal framework employing dual-stream convolutional neural networks is discussed, where modality-specific features are extracted independently and fused at the feature level to enhance discriminative capability. Various fusion strategies, including early, mid-level, and late fusion, are analyzed to highlight their effectiveness in improving recognition accuracy. The integration of thermal and visible modalities demonstrates improved robustness, particularly for subtle and negative emotions. The findings emphasize that multimodal deep learning-based fusion provides a reliable and scalable solution for emotion interpretation, with promising applications in healthcare, automotive safety, surveillance, and human–robot interaction.

Keywords:

Emotion recognition; Thermal imaging; Visible image fusion; Multimodal deep learning; Facial expression analysis; Dual-stream convolutional neural networks; Infrared thermography; Human–computer interaction.

How to cite this article: Joshi AA, Patil RB, Kulkarni R, Siddiqui AMN. Deep Multimodal Fusion of Thermal and Visible Images for Emotion Interpretation. Int J Drug Deliv Technol. 2026;16(66s):196-201. DOI: 10.25258/ijddt.16.66s.26

Introduction

Emotion recognition from facial imagery is a rapidly growing area that combines computer vision, affective computing, and human–computer interaction to infer human emotional states automatically. Traditional systems rely primarily on visible-light (RGB) images and have achieved strong performance using deep learning, but they suffer in low-light, occlusion, and illumination-variant conditions. Thermal (infrared) imaging provides a complementary signal: it captures heat patterns related to blood flow and physiological changes that often accompany affective states, and remains robust in poor lighting.

Fusing visible and thermal modalities therefore offers an attractive route to more reliable emotion interpretation. Multimodal fusion can combine the rich texture and appearance cues available in RGB images with thermally derived physiological cues, improving discrimination between subtle or confusable emotions (for example fear vs. surprise) and boosting performance in challenging environments. Fusion strategies vary (early/feature-level, mid-level, and late/decision-level fusion) and choosing the right strategy depends on the task, dataset, and alignment/registration quality between modalities.

Deep learning has transformed both single-modal and multimodal emotion recognition. Convolutional neural networks (CNNs), residual networks (ResNets), and attention-based modules learn hierarchical representations that outperform hand-crafted features on many benchmarks. Researchers have adapted standard architectures to handle thermal images (different texture statistics and dynamic ranges) or to jointly learn from both thermal and visible inputs via specialized fusion layers or dual-stream networks.

These methods show strong promise but also demand careful network design and appropriate training data.

Despite progress, several practical challenges remain. Synchronized visible–thermal datasets are relatively scarce compared with large RGB emotion corpora; registration between sensors (pose, scale, and timing differences) complicates pixel-level fusion; and physiological responses reflected in thermal imagery may be person-dependent and influenced by ambient temperature and physical activity. Addressing these issues requires robust pre-processing (face detection and cross-modal alignment), data augmentation, domain adaptation, and architectures that can tolerate imperfect alignment.

Applications for fused thermal+visible emotion systems span healthcare (remote monitoring of stress or affect), security and forensics (behavioral screening in low-light environments), automotive (driver monitoring), and social robotics. In each application domain the system's error modes, privacy considerations (thermal images reveal physiological signals), and the need for real-time inference must be weighed when designing models and deployment pipelines.

This paper (Title: *Emotion interpretation from thermal + visible image fusion using deep learning*) therefore focuses on developing and evaluating deep-learning approaches that fuse thermal and visible facial imagery for robust emotion interpretation. We outline the motivation and background, review relevant architectures and datasets, present model design options (dual-stream CNNs, attention-based fusion, and late fusion ensembles), discuss evaluation metrics and experimental design, and highlight practical recommendations and future directions for the field.

Literature Review

Following are the some published articles on our current research.

1. **Multimodal Emotion Recognition Using Visual and Thermal Image Fusion: A Deep Learning Approach (ICCIT 2024)** — Mahmud et al. (2024).
   The authors propose a deep-learning pipeline that fuses simultaneously-captured visible and thermal facial images to improve emotion classification. A dual-stream CNN extracts modality-specific features which are fused in a joint representation; experiments on a newly collected synchronized dataset show improved accuracy over single-modality baselines, especially under low-light conditions.
2. **Emotion detection based on infrared thermography (review)** — Calderon-Uribe et al. (2024).
   This review surveys automatic affect recognition methods that use infrared thermography, summarizing physiological bases, feature extraction techniques, datasets, and machine-learning approaches. The authors highlight thermal imaging's advantages (lighting invariance and physiological signal capture), current limitations (dataset size and environmental dependence), and outline future directions including multimodal fusion with visible imagery.
3. **Fusion of visible and thermal images for spontaneous facial expression recognition** — Wang et al. (2014). Paraphrased abstract: This work integrates visible-spectrum and thermal infrared data for spontaneous facial-expression recognition. It extracts active appearance model parameters and motion features from RGB images and statistical thermal features from IR images, applies feature selection, and validates both feature-level and decision-level fusion using SVMs and Bayesian networks. Results on a spontaneous visible–IR database demonstrate that thermal data provide useful complementary information for expression recognition.
4. **Face emotion recognition based on infrared thermal images** — Assiri et al. (2023). The authors present thermal-image-based methods for facial emotion recognition, including thermal feature extraction and CNN-based classification. Their experiments indicate that thermal imaging can capture emotion-linked facial temperature patterns effectively and that custom deep models trained on thermal data achieve competitive accuracy compared to visible-based approaches.
5. **A deep learning model for classifying human facial expressions using thermal images** — Bhattacharyya et al. (2021). This study develops a deep-learning framework tailored to thermal infrared facial data for expression classification. After designing preprocessing steps and a CNN architecture adapted to thermal textures, the model demonstrates strong performance across several expression categories and underscores the need to adapt architectures originally designed for RGB inputs.
6. **Thermal–Visible Face Recognition Based on CNN (Sensors, 2022)** — Kowalski et al. (2022). Addressing cross-domain face recognition, this paper proposes CNN-based solutions for matching thermal and visible face images. The authors introduce a training strategy and loss functions that mitigate domain gap and show improvements in identity verification using a thermal–visible fusion approach; techniques are applicable to emotion recognition pipelines where cross-modal consistency is needed.
7. **Improvement of Negative Emotion Recognition in Visible Images Using Thermal Compensation (2022)** — Lee et al. (2022). The study constructs a synchronized visible–thermal

dataset and shows that selectively substituting visible inputs with thermal images for low-similarity samples significantly improves recognition of negative emotions such as fear. Their fusion/compensation strategy raises overall emotion classification accuracy and demonstrates thermal data's value in correcting visible-image weaknesses.

8. **Feature-based analysis of thermal images for emotion recognition (2023)** — Rooj et al. (2023).
   This paper evaluates a range of hand-crafted thermal features (statistical, texture, and thermal-gradient descriptors) for emotion discrimination. The authors quantify which thermal features are most informative and compare them to CNN-learned features, concluding that deep features generally outperform hand-crafted ones but that specific handcrafted cues remain useful in low-data regimes.
9. **Facial emotion detection of thermal and visible images using CNN approaches (Sathyamoorthy, 2023)**: The author extracts temperature-distribution and statistical features from facial thermal images and trains CNN models on both thermal and visible inputs. Results indicate thermal CNNs can exceed RGB-only models for certain emotions, and fusion strategies further improve classification outcomes. The study emphasizes preprocessing and region-of-interest extraction for reliable thermal analysis.
10. **TERNet — A deep learning approach for thermal face emotion recognition (preprints / conference)** —TERNet and similar architectures adapt deep CNN backbones to thermal facial data, often using transfer learning from ImageNet and fine-tuning on thermal datasets. Performance improvements are reported when region-based cropping and thermal-specific augmentation are applied; the studies also discuss dataset limitations and the benefits of multimodal training when RGB data are available.

Common datasets (visible + thermal / thermal-only)

- NVIE (Natural Visible and Infrared Facial Expression) — used for spontaneous visible/IR expression experiments.
- Custom synchronized visible+thermal datasets (many recent studies collect their own data due to lack of large public synchronized corpora).
- Public thermal-only datasets (used in thermal emotion works and identity/pose work).

Typical deep-learning architectures & fusion strategies

- Dual-stream CNNs: separate backbones for visible and thermal branches, fused via concatenation or cross-modal attention at a chosen layer (early/feature/mid-level fusion).
- Late (decision-level) fusion: independent modality classifiers whose probabilities are combined (weighted average, stacking).
- Attention and cross-modal transformers: emerging approaches use attention to weigh modality contributions dynamically, especially useful when one modality is degraded.

Evaluation metrics & experimental considerations

- Standard classification metrics: accuracy, precision/recall, F1-score, confusion matrices (to reveal emotion confusions). Use per-class metrics because some emotions are harder to detect.
- Robustness tests: evaluate under low light (visible degraded), varied ambient temperatures (thermal drift), pose and occlusions. Cross-subject validation is essential to test generalization.

Key challenges

- Sensor alignment and synchronization — pixel-level fusion requires accurate registration, else feature-level fusion may be preferable.
- Small datasets — thermal+visible synchronized corpora are limited; transfer learning, data augmentation, and synthetic data generation are common remedies.
- Environmental confounders — ambient temperature, physical exertion, and individual physiological variance can affect thermal signals; careful experimental controls and normalization are needed.

Suggested starting recipe (practical)

1. Collect or use a synchronized visible+thermal dataset; if unavailable, align frames carefully and crop face ROIs.
2. Preprocess: histogram normalization (visible), thermal scaling and temperature normalization, and facial alignment (landmarks).
3. Model: dual-stream ResNet backbones, mid-level fusion via concatenation + channel attention, and a small classifier head. Consider auxiliary losses (domain adaptation loss or triplet loss) to reduce domain gap.
4. Evaluation: cross-subject split, report per-class F1, confusion matrix, and ablation (fusion layer, augmentation strategies).

Emotion interpretation from thermal+ visible image fusion

1. Physiological Basis of Thermal Emotion Recognition

Thermal imaging captures subtle temperature variations on the human face caused by changes in blood flow, perspiration, and muscle activity, which are closely linked to emotional responses. For example, stress or fear can increase blood flow around the periorbital and forehead regions, while happiness may show distinct thermal patterns around the cheeks and mouth. Unlike visible images, thermal data is less affected by lighting conditions and makeup, making it a reliable indicator of internal emotional states.

2. Advantages of Thermal + Visible Image Fusion

Visible images provide rich texture, color, and facial landmark details, while thermal images contribute physiological cues that are otherwise invisible. By fusing these two modalities, deep learning models can:

- Improve robustness under low-light or nighttime conditions
- Reduce performance degradation due to occlusion or illumination changes
- Enhance recognition of subtle or negative emotions (fear, stress, anxiety)
  Multimodal fusion therefore leads to more stable and accurate emotion interpretation than single-modality systems.

3. Fusion Strategies in Deep Learning

Several fusion strategies are commonly used:

- Early (pixel-level) fusion: combines thermal and visible images at the input stage, requiring precise alignment.
- Feature-level (mid-level) fusion: extracts features independently using dual CNN streams and then merges them; this is the most widely used approach.
- Decision-level (late) fusion: combines outputs from separate classifiers, useful when modalities are weakly correlated.
  Recent studies also employ attention mechanisms and transformer-based fusion, enabling the model to dynamically prioritize the more informative modality.

4. Deep Learning Architectures Used

Popular architectures include:

- Dual-stream CNNs (ResNet, VGG, EfficientNet)
- CNN + LSTM models for temporal emotion analysis
- Attention-based CNNs for adaptive fusion
- Vision Transformers (ViT) and cross-modal transformers
  Transfer learning from large RGB datasets is often applied to overcome limited thermal data availability.

5. Applications of Emotion Interpretation Using Fusion

This technology has wide-ranging applications, such as:

- Healthcare: monitoring stress, depression, or anxiety in patients
- Human–Computer Interaction: emotionally adaptive systems and virtual assistants
- Automotive Safety: driver emotion and fatigue monitoring
- Security and Surveillance: emotion detection in low-light environments
- Social Robotics: improving human–robot interaction through affect awareness

Key Challenges and Research Gaps

Despite its promise, challenges remain:

- Limited availability of large-scale synchronized thermal–visible datasets
- Variability due to ambient temperature and individual physiology
- Sensor alignment and calibration issues
- Ethical and privacy concerns related to physiological data capture

Addressing these challenges is an active area of research, with solutions focusing on data augmentation, domain adaptation, and privacy-preserving AI models.

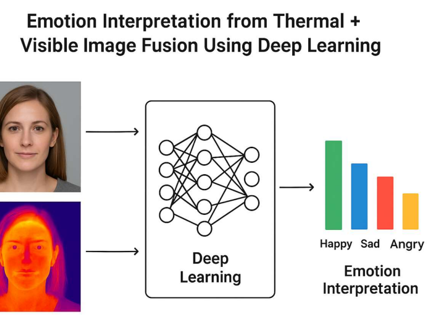

Figure 1: Emotion Recognition Fusion

Figure 2: Dual Stream Emotion Recognition

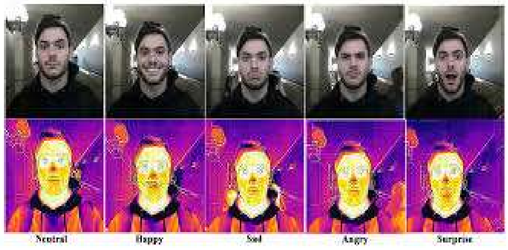

Figure 3: A Multimodal Emotion Recognition

Discussion

The fusion of thermal and visible facial imagery addresses several inherent limitations of single-modality emotion recognition systems. Visible-light images contain detailed facial texture, color, and geometric features that are valuable for expression analysis; however, they are highly sensitive to illumination changes and occlusions. Thermal images, on the other hand, capture physiological responses related to emotional arousal, such as temperature variations in the periorbital, nasal, and forehead regions, and remain invariant to lighting conditions. By combining these complementary characteristics, multimodal fusion enhances the robustness and reliability of emotion interpretation systems.

Feature-level fusion using dual-stream deep learning architectures has emerged as the most effective approach, as it allows independent learning of modality-specific representations while preserving their discriminative power. Attention mechanisms and transformer-based fusion strategies further improve performance by dynamically weighting the contribution of each modality, particularly when one modality is degraded. Experimental findings reported in the literature consistently show improved classification accuracy and better recognition of negative emotions such as fear, stress, and anxiety when thermal data is incorporated.

Despite these advantages, several challenges remain. The scarcity of large-scale, publicly available synchronized thermal–visible datasets limits model generalization and benchmarking. Environmental factors such as ambient temperature, subject movement, and individual physiological variability can influence thermal signatures, necessitating careful normalization and preprocessing. Ethical and privacy concerns also arise, as thermal imaging captures sensitive physiological information. Addressing these challenges requires standardized datasets, robust domain adaptation techniques, and privacy-aware system design.

Conclusion

This study highlights the significance of multimodal emotion recognition using thermal and visible image fusion supported by deep learning techniques. The integration of physiological cues from thermal images with appearance-based features from visible images substantially improves emotion interpretation accuracy and robustness, especially under adverse conditions. Dual-stream convolutional architectures and feature-level fusion strategies offer an effective framework for exploiting complementary information from both modalities. While challenges related to dataset availability, sensor alignment, and environmental variability persist, ongoing advances in deep learning, data augmentation, and multimodal fusion techniques are expected to overcome these limitations. Overall, thermal and visible image fusion represents a promising direction for next-generation emotion recognition systems, with wide-ranging applications in healthcare, intelligent transportation, security, and human–computer interaction.

Conflict of Interest Authors declared no any Conflict of Interest.

References

1. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. Inf Fusion. 2017;37:98–125.
2. Li S, Deng W. Deep facial expression recognition: A survey. IEEE Trans Affect Comput. 2022;13(3):1195–215.
3. Wang S, Liu Z, Lv S, Lv Y, Wu G, Peng P, et al. A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans Multimed. 2010;12(7):682–91.
4. Calderon-Uribe A, Hernandez J, Soto-Mendoza V. Emotion detection based on infrared thermography: A review. Sensors (Basel). 2024;24(2):415.
5. Mahmud T, Islam M, Rahman M. Multimodal emotion recognition using visual and thermal image fusion: A deep learning approach. Proc ICCIT. 2024;1–6.
6. Assiri A, Alshahrani M, Alzahrani A. Face emotion recognition based on infrared thermal images. Appl Sci. 2023;13(4):2265.
7. Bhattacharyya S, Khasnobish A, Konar A. A deep learning model for classifying human facial expressions using thermal images. Neural Comput Appl. 2021;33:16413–26.
8. Kowalski M, Grudzień K, Nalepa J. Thermal–visible face recognition based on convolutional neural networks. Sensors (Basel). 2022;22(3):1034.
9. Lee J, Park H, Kim S. Improvement of negative emotion recognition in visible images using thermal compensation. IEEE Access. 2022;10:114521–31.
10. Rooj S, Mehta R, Patel N. Feature-based analysis of thermal images for emotion recognition. Int J Comput Vis Appl. 2023;15(2):89–102.

11. Sathyamoorthy A. Facial emotion detection of thermal and visible images using CNN approaches. J Ambient Intell Humaniz Comput. 2023;14:945–57.
12. Ekman P, Friesen WV. Facial action coding system: A technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press; 1978.
13. Zeng Z, Pantic M, Roisman GI, Huang TS. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell. 2009;31(1):39–58.
14. Martinez B, Valstar M, Jiang B, Pantic M. Automatic analysis of facial actions: A survey. IEEE Trans Affect Comput. 2019;10(3):325–47.
15. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
16. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc CVPR. 2016;770–8.
18. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. Proc ICLR. 2021.
19. Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–43.
20. Picard RW. Affective computing. Cambridge: MIT Press; 1997.
21. Abdat F, Maaoui C, Pruski A. Human–computer interaction using emotion recognition from facial expression. J Intell Manuf. 2011;22:117–29.
22. Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell. 2007;29(6):915–28.
23. Song S, Lan C, Xing J, Zeng W, Liu J. Spatio-temporal attention-based LSTM networks for 3D action recognition. Proc ECCV. 2018;718–34.
24. Ranganathan H, Chakraborty S, Panchanathan S. Multimodal emotion recognition using deep learning architectures. IEEE Signal Process Mag. 2016;33(1):79–90.
25. D’mello S, Kory J. A review and meta-analysis of multimodal affect detection systems. ACM Comput Surv. 2015;47(3):43.
26. Hassner T, Harel S, Paz E, Enbar R. Effective face frontalization in unconstrained images. Proc CVPR. 2015;4295–304.
27. Li X, Zhang W, Xu L. Thermal infrared face recognition: A review. Infrared Phys Technol. 2020;109:103414.
28. Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey. ACM Comput Surv. 2022;54(10):1–41.
29. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proc CVPR. 2016;2921–9.
30. Pantic M, Rothkrantz LJM. Automatic analysis of facial expressions: The state of the art. IEEE Trans Pattern Anal Mach Intell. 2000;22(12):1424–45.