Vol. 4, No. 1, pp.: 1-5, 2026

Multimodal Fusion Strategies for Affective Computing: Audio, Visual, and Physiological Signals

Mei-Ling Zhou[1,*], **Kwame Asante**[2], **Irina Petrov**[3]

[1] Brain and Cognitive Science Department, MIT, Cambridge, MA, USA, 02139

[2] Department of Electrical Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, AK385

[3] Higher School of Economics, Faculty of Computer Science, Moscow, Russia, 101000

[*] *meizhou@mit.edu*

Article Information	
Received	20 November 2025
Accepted	30 March 2026
DOI	https://doi.org/10.63646/datamind.2026.040101

Abstract

Affective computing — the computational recognition, modelling, and response to human emotion — has been a research area for nearly three decades, but the convergence of multimodal deep learning, wearable sensor technology, and large-scale pre-trained models has opened new possibilities and surfaced old problems in sharper relief. This review surveys multimodal fusion strategies for affective computing, focusing on the combination of three signal modalities that together cover the principal pathways through which emotion is expressed: audio (speech prosody, paralinguistic features), visual (facial action units, body posture, gaze), and physiological signals (EEG, EDA, heart rate variability). We analyse fusion architectures across four families — feature-level, decision-level, model-level, and hybrid — and evaluate their performance on six benchmark datasets spanning three affective recognition tasks: valence-arousal regression, discrete emotion classification, and pain intensity estimation. A key finding is that physiological signal modalities are systematically underrepresented in multimodal fusion research relative to their information value, partly due to data collection constraints and partly due to cultural assumptions about which emotion expressions are universal versus culturally specific. We identify cross-cultural generalisation and privacy-preserving physiological sensing as the most critical open research directions.

Keywords: ***affective computing; multimodal fusion; emotion recognition; EEG; audio-visual; physiological signals; cross-cultural***

1. Introduction

The aspiration to build machines that can recognise and respond to human emotion is almost as old as artificial intelligence itself. Rosalind Picard's foundational 1997 book gave the field its name and a coherent research agenda that has proved remarkably durable. What has changed since 1997 is not the aspiration but the available tools: deep neural networks that can extract rich representations from speech and face images without hand-crafted features, wearable sensors that bring physiological measurement out of the laboratory and into everyday life, and pre-trained transformer models that encode rich contextual understanding of language and paralinguistic signals.

These advances have produced measurable progress on the standard benchmarks. But they have also sharpened the visibility of some fundamental problems that the field has not fully resolved. How do you fuse modalities that have different temporal resolutions, different noise characteristics, and different reliabilities across individuals and contexts? How do you build affective recognition systems that generalise across cultures when the available training data is dominated by Western, WEIRD (White, Educated, Industrialised, Rich, Democratic) populations? And what are the appropriate ethical constraints on deploying systems that can infer people's emotional states from passive physiological monitoring?

2. Modality Overview

2.1 Audio Modalities

Speech carries emotion in two largely separable channels: the linguistic content (what is said) and the paralinguistic features (how it is said — prosody, rhythm, voice quality). For automatic speech recognition-based emotion analysis, the linguistic content is increasingly well-handled by pre-trained language models. The paralinguistic dimension, captured through features like fundamental frequency contour, speech rate, and formant trajectories, requires either explicit acoustic feature extraction (openSMILE, GeMAPS) or end-to-end learning from waveforms (wav2vec 2.0, WavLM fine-tuned on affective tasks).

The challenge for audio emotion recognition is speaker normalisation: the same acoustic feature values mean very different things for different speakers. A monotone delivery that signals boredom in one speaker may be a baseline characteristic of another's speech style. Appropriate normalisation — either through speaker-dependent models or through self-normalisation against a speaker-specific baseline — remains an active research challenge.

2.2 Visual Modalities

Facial expression analysis, the visual modality most directly associated with emotion recognition, has a well-established anatomical grounding in the Facial Action Coding System (FACS) developed by Ekman and Friesen. Automated FACS-based analysis identifies the activation of individual facial muscles (Action Units) and maps them to discrete emotion categories or dimensional valence-arousal coordinates. Modern deep learning approaches to facial emotion recognition primarily use CNN or ViT backbones trained on labelled datasets like AffectNet (450K images) and RAF-DB (30K images). Performance on controlled lab datasets is high; performance drops significantly in unconstrained, in-the-wild settings due to occlusion, head pose variation, and image quality.

2.3 Physiological Modalities

Physiological signals — electroencephalography (EEG), electrodermal activity (EDA), heart rate variability (HRV), skin temperature, and electromyography (EMG) — provide direct measures of the autonomic and central nervous system responses that accompany emotional experience. They are, in principle, harder to voluntarily suppress or falsify than facial expressions or speech prosody, which makes them particularly valuable for applications where posed or masked expressions are expected. Their practical disadvantage is measurement complexity: EEG requires careful electrode placement and is sensitive to motion artifact; EDA and HRV require wearable sensors that some populations find intrusive.

3. Fusion Architectures

Table 1. *Comparison of multimodal fusion families across five dimensions for affective computing applications.*

Fusion Family	Fusion Stage	Modality Alignment Needed	Handles Missing Modality	Interpretability	Best Use Case
Feature-level	Early	Yes (temporal)	Poorly	Low	Controlled lab, synchronised capture
Decision-level	Late	No	Well	High	Heterogeneous sensors, unequal reliability
Model-level	Middle	Partial	Moderate	Medium	General-purpose, sufficient data
Hybrid	Multiple	Partial	Moderate	Medium	High-performance, research settings

'Modality alignment needed' refers to temporal synchronisation requirement. Missing modality handling is rated relative to other fusion types in the same evaluation setting.

3.1 Feature-Level (Early) Fusion

Feature-level fusion concatenates representations from multiple modalities before the classification or regression head. The appeal is its simplicity: a single model learns to weight all modality features jointly. The constraint is temporal alignment — if the audio features are extracted at 25fps and the physiological signal is sampled at 256Hz, they must be resampled or windowed to a common temporal resolution before concatenation. Early fusion is most appropriate when modalities are tightly synchronised in time and the sensor configuration is controlled.

3.2 Decision-Level (Late) Fusion

Decision-level fusion trains modality-specific models independently and combines their outputs (scores, probabilities, or predicted labels) using a fusion rule or a learned combiner. The advantages are robustness to missing modalities, the ability to use different architectures optimised for each modality, and interpretability — you can inspect each modality's contribution. Late fusion is particularly well-suited for applications where modality availability varies at inference time.

3.3 Model-Level (Intermediate) Fusion

Intermediate fusion — typically implemented through cross-attention mechanisms between modality-specific encoders — has become the dominant approach in high-performance affective computing systems following the success of transformer-based architectures. The MulT model (Tsai et al., 2019) introduced directional cross-modal attention that allows each modality's representation to attend to the others at multiple time steps. This approach captures cross-modal dependencies that neither early nor late fusion can represent, at the cost of requiring aligned, synchronised inputs.

[Figure 1 — Performance comparison of fusion families on valence-arousal regression (CCC metric) across six affective computing benchmark datasets (IEMOCAP, AVEC16, MAHNOB-HCI, DEAP, CMU-MOSI, CMU-MOSEI). Intermediate and hybrid fusion consistently outperform early and late fusion; the margin is largest on datasets with physiological modalities (DEAP, MAHNOB-HCI). Error bars represent ± 1 SD across three random seeds.]

Figure 1. *Performance comparison of fusion families on valence-arousal regression (CCC metric) across*

six affective computing benchmark datasets (IEMOCAP, AVEC16, MAHNOB-HCI, DEAP, CMU-MOSI, CMU-MOSEI). Intermediate and hybrid fusion consistently outperform early and late fusion; the margin is largest on datasets with physiological modalities (DEAP, MAHNOB-HCI). Error bars represent ± 1 SD across three random seeds.

4. Cross-Cultural Generalisation

The cross-cultural limitation of affective computing systems is among the field's most significant and least discussed problems. Ekman's universality thesis — that basic emotion expressions are innate and cross-culturally consistent — has been the implicit theoretical foundation of much affective computing research, justifying the transfer of models trained predominantly on North American and European populations to global deployment. The empirical support for strong universality has been challenged significantly by cross-cultural psychology research over the past two decades. Barrett and colleagues have shown that emotion category boundaries vary substantially across cultures; cultural display rules govern when and how emotions are expressed publicly; and the dimensional (valence-arousal) space that underlies most computational models may itself be culturally constructed.

For affective computing, these findings imply that a model trained on IEMOCAP (American English, predominantly non-Hispanic white actors) will systematically misclassify emotional states in speakers from other cultural and linguistic backgrounds. The available cross-cultural affective datasets are insufficient to redress this gap: CHEAVD (Chinese), MEC (Chinese), and BAUM-2s (Turkish) each cover a small fraction of the world's cultural and linguistic diversity. Building truly cross-cultural affective computing systems will require sustained investment in diverse, ecologically valid data collection — a resource-intensive effort that the current academic incentive structure does not naturally encourage.

5. Ethical Considerations

We want to state plainly what this review's survey of emotional signal fusion implies for deployment contexts. Emotion recognition systems applied to surveillance, hiring, or forensic assessment contexts — all of which have been commercially deployed — are ethically concerning for reasons that do not disappear when the underlying technology improves. The concern is not primarily about current accuracy limitations, though those are real. It is about the epistemic authority granted to computational systems that infer private mental states from observable signals. Even a perfectly accurate emotion classifier — which does not and cannot exist — raises profound questions about consent, privacy, and the appropriate uses of emotional inference in contexts where power relationships are unequal. The affective computing community has a responsibility to engage with these questions directly rather than deferring them to future work.

6. Conclusion

Multimodal fusion has produced the best-performing affective computing systems to date, and the integration of physiological signals represents the most promising near-term frontier for improving both accuracy and robustness. The field faces two challenges that technical progress alone cannot resolve: the cultural specificity of available training data and the ethical dimensions of deploying emotional inference systems. These challenges deserve the same creative energy and resource investment that the fusion architecture problem has received. Progress on the technical frontier without equivalent progress on the cultural and ethical frontiers will produce systems that are more capable but not more trustworthy.

References

Picard, R. W. (1997). Affective computing. MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. Consulting Psychologists Press. https://doi.org/10.1037/t27734-000

Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. ACL 2019, 6558–6569. https://doi.org/10.18653/v1/P19-1656

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., & Pantic, M. (2017). AVEC 2017: Real-life depression, and affect recognition workshop and challenge. AVEC@ACM MM 2017, 3–9. https://doi.org/10.1145/3133944.3133953

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., & Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. ACL 2018, 2236–2246. https://doi.org/10.18653/v1/P18-1208

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychological Science in the Public Interest, 20(1), 1–68. https://doi.org/10.1177/1529100619832930

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. Interspeech 2009, 312–315. https://www.isca-speech.org/archive/interspeech_2009/schuller09_interspeech.html

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923