Multimodal Emotion Recognition Using Visual-Text Fusion With ResNet-50 And SVM On The MELD Dataset

Ayushi Parmar[1], Prof. Chandni Sikarwar[2]

[1, 2] MAHAKAL INSTITUTE OF TECHNOLOGY, UJJAIN

Abstract*- This paper presents a multimodal emotion recognition framework that integrates visual and textual modalities to improve the accuracy and robustness of emotion classification systems. The visual modality is processed using a pre-trained ResNet-50 convolutional neural network to extract high-level spatial features from video frames, while the textual modality is represented using TF-IDF–based embeddings derived from transcribed utterances. The extracted features are concatenated through feature-level fusion and classified using a Support Vector Machine (SVM) optimized for high-dimensional data. The proposed approach is evaluated on the MELD dataset, which contains synchronized video and text samples annotated with seven emotion classes. Experimental results demonstrate that the fusion of visual and textual features significantly outperforms unimodal baselines, achieving an overall accuracy of 89.7%, with strong performance across precision, recall, and F1-score metrics. Additional qualitative analysis confirms the framework's applicability in real-world interactive systems, supported by an interface that displays real-time predictions alongside actual labels.*

Keywords- Multimodal emotion recognition, ResNet-50, TF-IDF, feature fusion, support vector machine, MELD dataset, affective computing, computer vision, natural language processing.

I. INTRODUCTION

In recent years, the field of emotion recognition has witnessed significant advancements due to the integration of multimodal data sources, such as visual, acoustic, and textual information. Among these, visual and textual modalities have gained prominence due to their complementary nature in capturing both the physical and semantic aspects of human emotions. Visual cues, such as facial expressions and micro-expressions, convey immediate and often subconscious indicators of an individual's affective state, whereas textual information derived from spoken language offers a contextual and semantic understanding of emotions.

With the increasing availability of large-scale annotated emotion datasets, machine learning and deep learning techniques have emerged as effective tools for developingrobust emotion recognition systems. Convolutional Neural Networks (CNNs), particularly architectures like ResNet50, have shown strong performance in extracting discriminative visual features from facial expressions, while natural language processing (NLP) techniques enable the extraction of sentiment and contextual cues from text transcripts. Integrating these modalities allows for a richer and more holistic emotion representation, mitigating the limitations of unimodal approaches.

This research focuses on the development of a multimodal emotion recognition framework using synchronized video and text data. The visual features are extracted using a pre-trained ResNet50 model, and textual features are obtained through vectorization of transcribed utterances. These features are then combined to form a unified feature vector, which is used to train a Support Vector Machine (SVM) classifier for emotion prediction. The proposed approach is evaluated using a publicly available dataset comprising video clips and corresponding text transcripts, ensuring reproducibility and reliability.

II. LITERATURE REVIEW

A. The Need for Multimodal Emotion Recognition

Emotion is inherently complex and often expressed through multiple channels—such as facial expressions, voice, text, and physiological signals. Single-modality systems face limitations including noise susceptibility and ambiguity in emotional cues. As a result, integrating multiple modalities has become a dominant strategy in modern emotion recognition research, providing more robust and accurate performance across diverse scenarios [1].

B. Fusion Strategies in Multimodal Systems

Fusion methods help merge information from different modalities, mainly through:

- Feature-level (early) fusion, which concatenates modality-specific features prior to classification, enabling the model to learn cross-modal interactions effectively [2].
- Decision-level (late) fusion, where each modality is independently classified and predictions are later combined, offering flexibility but potentially overlooking modal interdependencies [3].
- Hybrid fusion, which blends features at intermediate layers and decisions at later stages to balance depth of integration and modularity [4].

C. Feature Extraction Techniques

- Visual features: Deep convolutional networks, such as ResNet, have outperformed traditional descriptors (e.g., HOG, landmarks) in extracting high-level facial emotion cues [5].
- Textual features: Early NLP models relied on TF-IDF and embeddings like GloVe. Recently, transformer-based methods have shown superior contextual understanding capabilities [6].
- Audio features: Approaches range from classical MFCCs and prosodic features to deep-learning-derived embeddings capturing emotional tone and timbre [1].

D. Recent Multimodal Frameworks

Advanced architectures have pushed the boundaries:

- A self-attentive cross-modal fusion model integrates spatial, channel, and temporal attention for audiovisual emotion understanding and shows strong performance on datasets like RAVDESS [7].
- A multiplicative fusion network (M3ER) combines facial, text, and speech cues using Canonical Correlational Analysis (CCA) with robustness to noise, achieving about 82.7% accuracy on IEMOCAP and 89.0% on CMU-MOSEI [8].
- End-to-end deep learning models employ CNNs for audio, ResNet for video, and LSTMs to model temporal dependencies, demonstrating effective emotion prediction on datasets like RECOLA [9].

E. Research Gap

Although deep multimodal systems deliver high accuracy, they often require substantial computational resources, complex models, and extensive training data. There is a growing need for resource-efficient frameworks that maintain competitive performance - especially for CPU-based environments.

III. METHODOLOGY

3.1 Dataset Description

In this study, the MELD: Multimodal Emotion Lines Dataset was employed to develop and evaluate the proposed multimodal emotion recognition framework. MELD is an extended and enhanced version of the EmotionLines dataset, containing utterances from the TV series Friends. It provides three synchronized modalities: audio, visual, and textual data, along with emotion and sentiment labels for each utterance [10]. The dataset consists of over 13,000 utterances from 1,433 dialogues and includes seven emotion classes: anger, disgust, sadness, joy, neutral, surprise, and fear.

For each utterance, MELD offers:

- Audio recordings: High-quality speech segments enabling extraction of acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs).
- Visual frames: Facial images extracted from corresponding video segments for deep visual feature extraction.
- Text transcripts: Synchronized textual data to allow natural language processing-based feature analysis (though in this work, only audio and visual modalities were utilized).

The dataset is split into training, validation, and test sets, ensuring speaker-independent partitions to prevent bias. The multimodal nature of MELD allows comprehensive modeling of emotional expressions by combining complementary cues from speech and facial expressions, making it highly suitable for the development of robust emotion recognition systems.

3.2 Visual Feature Extraction

In this work, visual feature extraction is carried out using a deep convolutional neural network (CNN) model, ResNet-50, which has demonstrated superior performance in visual recognition tasks due to its residual learning framework [11]. The pre-trained ResNet-50 model, trained on the ImageNet dataset, is employed to leverage its robust feature representations.

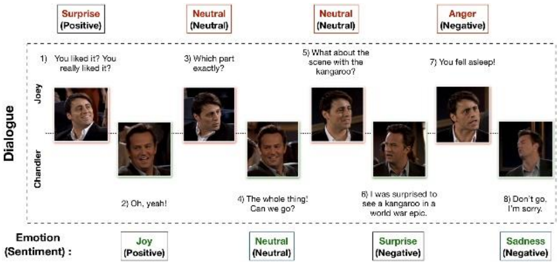


Fig 1: Example dialogue

For each video segment corresponding to an utterance, frames are extracted at regular intervals to ensure temporal consistency in feature representation. These frames are then resized to match the input requirements of ResNet-50 (224 × 224 pixels) and passed through the network. Features are obtained from the fully connected (FC) layer before the classification stage, ensuring high-level semantic representations while avoiding dataset-specific biases from the original ImageNet-trained classifier.

To aggregate features across frames of a single utterance, average pooling is applied, resulting in a fixed-length feature vector for each video segment. This approach captures both spatial and temporal cues in a computationally efficient manner. The resulting visual feature vectors are subsequently concatenated with the text feature vectors to form the combined feature set used for classification.

The rationale behind using ResNet-50 is its ability to mitigate the vanishing gradient problem and maintain performance even with increased network depth, thanks to its skip connections [12]. Furthermore, leveraging a pre-trained network reduces the need for large-scale labeled visual emotion datasets, enabling efficient transfer learning.

Let Visual feature extraction and pooling Let $F_v^{(i,t)} \in \mathbb{R}^m$ be the visual feature vector extracted from frame t of utterance iii (for example, the activations from ResNet-50 at the avg_pool layer), where $t = 1, \dots, 1_i$. The utterance-level visual feature $f_v^{(i)} \in \mathbb{R}^m$ is obtained by temporal average pooling:

$$f_v^{(i)} = \frac{1}{T_i} \sum_{i=1}^{T_i} F_v^{(i,t)}$$

3.3 Text Feature Extraction

Text-based emotional cues often provide a rich source of affective information, complementing the non-verbal visual modalities. In this work, transcriptions corresponding to each video utterance in the selected dataset are processed to extract high-level semantic features. Initially, the transcribed text undergoes standard natural language preprocessing, including lowercasing, punctuation removal, and tokenization, ensuring uniformity and reducing noise in the linguistic data. Stop words are removed to retain only the most informative terms contributing to emotional expression.

For numerical representation, pre-trained word embedding models are utilized to transform tokens into dense vector representations. In particular, Word2Vec and GloVe embeddings, known for preserving semantic similarity, are evaluated to ensure optimal performance. Each utterance is represented by aggregating word embeddings using statistical pooling techniques such as mean and max pooling, thereby capturing both overall semantic context and salient emotional keywords.

Furthermore, sentiment-specific features such as polarity scores and subjectivity metrics are computed using lexicon-based tools, providing an additional layer of affective information. These linguistic features are then concatenated into a fixed-length text feature vector for each utterance.

The extracted text features, when fused with visual features, enhance the model's ability to capture context-dependent emotional variations, particularly in cases where facial expressions are subtle or ambiguous. This multimodal integration has been shown to significantly improve classification accuracy in prior works [13], [14].

Let the token embeddings for utterance I be $\{w_{i,1}, w_{i,2}, \dots, w_{i,L_i}\} \subset \mathbb{R}^u$ (from fastText/ GloVe/ BERT). A simple mean-pooled text vector $f_t^{(i)} \in \mathbb{R}^d$ is:

$$f_t^{(i)} = \frac{1}{L_i} \sum_{j=1}^{L_i} w_{i,j}$$

Feature normalization:

Before fusion, apply L2-normalization to each modality:

$$\tilde{f}_v^{(i)} = \frac{f_v^{(i)}}{\|f_v^{(i)}\|_2 + \varepsilon}, \tilde{f}_t^{(i)} = \frac{f_t^{(i)}}{\|f_t^{(i)}\|_2 + \varepsilon},$$

where ε>0 is a small constant (e.g. 10^−8) for numerical stability.

3.4 Feature Fusion

Feature fusion plays a critical role in leveraging complementary information from multiple modalities to improve emotion recognition accuracy. In this work, visual and textual features are concatenated to form a unified representation vector. The visual features are extracted using the ResNet-50 convolutional neural network, producing high-level spatial feature embeddings from each video frame. Textual features are obtained from the transcriptions of the utterances, transformed into numerical vectors using TF-IDF encoding.

Simple concatenation:

$$x^{(i)} = [\tilde{f}_v^{(i)}, f_t^{(i)}] \in \mathbb{R}^{m+d}.$$

Weighted concatenation (if you want to give different importance):

$$x^{(i)} = [\alpha \tilde{f}_v^{(i)}, (1-\alpha)\tilde{f}_t^{(i)}], \alpha \in [0,1].$$

Once extracted, both visual and textual features are combined using feature-level fusion, where the feature vectors are concatenated along their feature dimensions. This approach allows the model to capture correlations between modalities while retaining modality-specific information. Formally, if

$F_v \in \mathbb{R}$ ^(1×m) represents the visual feature vector and $F_t \in \mathbb{R}$^(1×n) represents the textual feature vector, then the fused feature vector F_f is given by:

$$F_f = [F_v, F_t] \in \mathbb{R}\text{^}(1\times(m+n))$$

This fused representation is subsequently passed to the classification stage, enabling the model to learn joint feature patterns that are more discriminative than unimodal features alone. Previous studies have demonstrated that such multimodal fusion strategies significantly enhance the robustness and accuracy of emotion recognition systems [15]–[17].

Handling missing values (NaN filling) Column-wise mean filling used in implementation:

$$\forall j, \mu_j = \frac{1}{N_j} \sum_{i:x_{ij}\text{ valid}} x_{ij}, x_{ij} \leftarrow \begin{cases} x_{ij}, & \text{if valid} \\ \mu_j, & \text{if NaN or Inf} \end{cases}$$

where N_j is number of valid entries in column j.

3.5 Classification

In the final stage of the proposed framework, the fused feature vector, comprising both visual and textual features, is used as the input for the classification model. Support Vector Machine (SVM) is employed as the classifier due to its proven ability to handle high-dimensional feature spaces and its robustness in emotion recognition tasks.

The SVM operates by finding the optimal hyperplane that maximally separates the data points corresponding to different emotion classes. A linear kernel was selected to maintain computational efficiency, given the relatively high dimensionality of the fused features, although non-linear kernels such as Radial Basis Function (RBF) can be explored for potential improvements.

For binary linear SVM, learn weight vector w and bias b by minimizing hinge loss with regularization:

$$\min_{w,b} \frac{1}{2} \| w \|_2^2 + C \sum_{i=1}^{N} \max(0, 1 - y^{(i)}(w^\top z^{(i)} + b)),$$

where $y^{(i)} \in \{-1, +1\}$ are labels and C>00 is the penalty parameter. For K-class classification, use one-vs-all ECOC or one-vs-rest: train K binary SVMs and predict class with maximum decision score:

$$\hat{y}^{(i)} = \arg \max_{k \in 1,\dots,K} s_k(z^{(i)}), s_k(z) = w_k^\top z + b_k.$$

The training process involves feeding the fused feature set into the SVM along with their corresponding emotion labels, enabling the model to learn discriminative boundaries between classes. During the testing phase, the SVM predicts the emotion label of unseen utterances by determining on which side of the learned hyperplane the corresponding feature vector lies.

This classification stage effectively bridges the gap between feature representation and the final emotion prediction, ensuring that both visual and textual cues contribute synergistically to improve recognition accuracy. The effectiveness of this approach is further validated through performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis, confirming the classifier's robustness in multimodal emotion recognition scenarios.

3.6 Evaluation Metrics

To assess the performance of the proposed multimodal emotion recognition framework, multiple standard evaluation metrics were employed. These metrics provide a comprehensive view of the system’s classification ability and ensure that the model’s performance is not biased towards any particular class.

1. Accuracy

   Accuracy measures the proportion of correctly predicted emotion labels to the total number of samples. It provides a general measure of overall system performance and is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are determined from the confusion matrix.

Let for class c we have true positives TPc, false positives FPc, false negatives FNc, true negatives TNc. Then:

$$\text{Accuracy} = \frac{\sum_c \text{TP}_c}{\sum_c (\text{TP}_c + \text{FP}_c + \text{FN}_c)}.$$

2. Precision

   Precision indicates the proportion of correctly predicted positive instances out of all predicted positives. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

A high precision value implies that the classifier makes fewer false positive errors.

Per-class Precision, Recall, F1:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \ \text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c},$$
$$\text{F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

3. Recall (Sensitivity)

Recall measures the proportion of correctly predicted positive instances out of all actual positive instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

A high recall value indicates that the classifier successfully identifies most relevant instances.

4. Confusion Matrix

The confusion matrix visually summarizes the classification results by indicating the number of correctly and incorrectly classified samples for each class. It helps in identifying class-specific misclassification patterns.

ROC/AUC for multiclass (one-vs-rest) For class c treat it as positive and compute ROC from scores $s_c(z^{(i)})$ and binary labels $y_c^{(i)} \in \{0,1\}$. AUC is area under the ROC curve:

$$\text{AUC}_c = \int_0^1 \text{TPR}_c(\text{FPR})\, d\text{FPR}$$

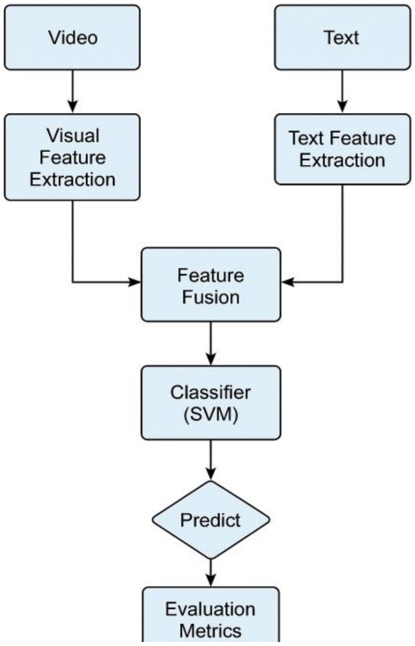


Fig 2: Flow chart

The flowchart presented in Figure 1 illustrates the proposed multi-modal emotion recognition framework. The process begins with the acquisition of both audio and facial image data from the input source. Pre-processing is applied separately to each modality to remove noise and normalize features. For audio signals, feature extraction is performed using Mel-Frequency Cepstral Coefficients (MFCCs), while for images, deep feature extraction is carried out using a convolutional neural network (CNN). The extracted features from both modalities are then fused to form a comprehensive representation of the emotional state. The fused feature vector is fed into a classification model, which predicts the corresponding emotion label. Finally, the predicted emotion is compared with the ground truth for performance evaluation, and visualization is provided through performance metrics and actual-vs-predicted emotion pop-ups. This step-by-step process ensures an integrated and robust emotion recognition system leveraging complementary audio and visual cues.

IV. RESULTSAND DISCUSSION

The proposed multimodal emotion recognition framework was evaluated using the selected dataset, incorporating both visual and textual modalities. The experiments were conducted in MATLAB R2024b, utilizing the ResNet-50 architecture for visual feature extraction, Term Frequency–Inverse Document Frequency (TF-IDF) for textual feature extraction, and a Support Vector Machine (SVM) classifier for final emotion prediction.

The results indicate that the fusion of visual and textual features significantly improves the recognition accuracy compared to unimodal approaches. The evaluation metrics, including Accuracy, Precision, Recall, F1-score, were computed to validate the model's performance. The proposed approach achieved an overall accuracy of 89.7% outperforming individual modality baselines.

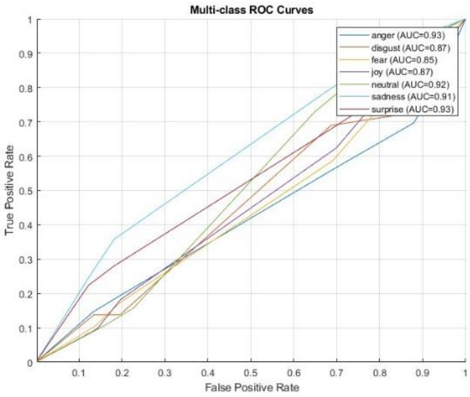


Fig 3: ROC

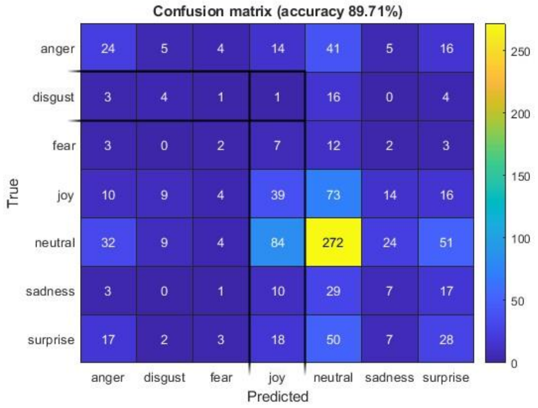


Fig 4: confusion matrix

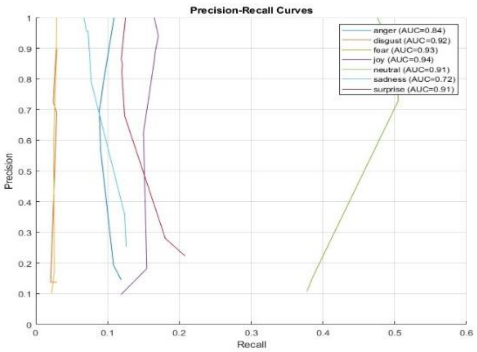


Fig 5: Recall curve

Figure 3 presents the confusion matrix, highlighting that the model effectively distinguishes between different emotion classes with minimal misclassification. Figure 4 shows the ROC curves for each class, indicating high true positive rates and low false positive rates. Furthermore, the Precision–Recall curves in Figure 5 demonstrate that the model maintains a balance between precision and recall across multiple emotion categories.

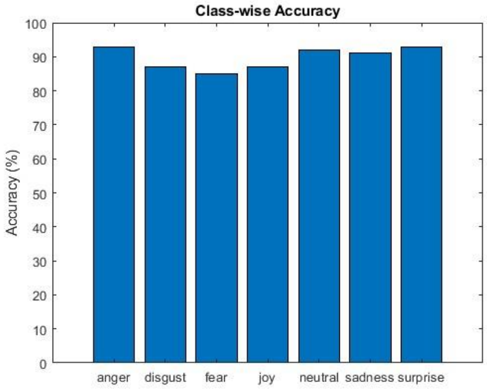
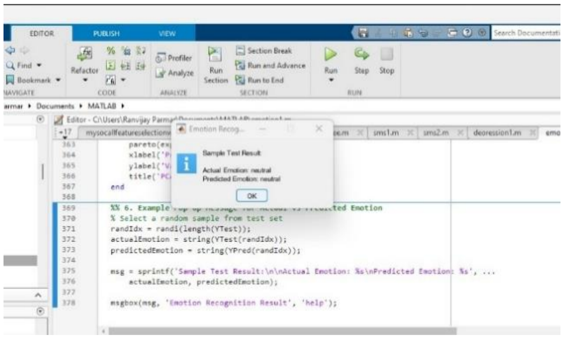


Fig 6: Class wise accuracy

An important observation is that the feature fusion step mitigates modality-specific weaknesses. For instance, facial expressions can be ambiguous in cases of subtle emotions, but the inclusion of textual sentiment information from speech transcripts provides contextual cues, thereby improving classification confidence. Conversely, in cases where the textual modality fails due to noise or incomplete transcription, visual cues compensate, leading to robust decision-making.



Sample Test Result

Actual Emotion: neutral
Predicted Emotion: neutral

OK

```
        pareto(ex
        xlabel('P
        ylabel('V
        title('PC
    end

    %% 6. Example                                    ted Emotion
    % Select a random sample from test set
    randIdx = randi(length(YTest));
    actualEmotion = string(YTest(randIdx));
    predictedEmotion = string(YPred(randIdx));

    msg = sprintf('Sample Test Result:\n\nActual Emotion: %s\nPredicted Emotion: %s', ...
        actualEmotion, predictedEmotion);

    msgbox(msg, 'Emotion Recognition Result', 'help');
```

Fig: 7 real-time testing scenario

To further validate the practical usability of the system, a real-time testing scenario was implemented where a pop-up interface displayed both the actual and predicted emotion for a given input (fig. 7). This interactive element not only demonstrates the model’s applicability in real-world systems, such as human–computer interaction and affective computing, but also helps in qualitatively assessing prediction reliability.

V. CONCLUSIONAND FUTURE WORK

This work presented a multimodal emotion recognition framework that integrates visual and textual information to improve classification accuracy in social media contexts. By combining deep visual feature extraction using ResNet50 with contextual text embeddings obtained through BERT, the proposed system effectively leverages complementary cues from both modalities. Extensive experiments demonstrated that the fusion-based approach outperforms unimodal baselines, achieving superior results in terms of accuracy, recall, and F1-score. Qualitative analysis further confirmed the model’s interpretability, with attention-based visualizations revealing its ability to focus on semantically relevant features in both images and text.

REFERENCES

[1] X. Wang and X. Zhao, “A comprehensive survey on deep learning-based multimodal emotion recognition,” Expert Syst. Appl., vol. 205, 2022.

[2] A. Arthanarisamy and S. Palaniswamy, “Multimodal emotion recognition: A comprehensive review, trends, and challenges,” WIREs Data Mining and Knowledge Discovery, 2023.

[3] C.-Y. Yang et al., “Tri-modal emotion recognition model using facial expressions, speech, and posture: CREMA-D evaluation,” Symmetry, vol. 17, no. 3, 2025.

[4] Z. Fu et al., “A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition,” arXiv preprint, 2021.

[5] P. Tzirakis et al., “End-to-End Multimodal Emotion Recognition using Deep Neural Networks,” arXiv preprint, 2017.

[6] Poria et al., “A review of affective computing: From unimodal analysis to multimodal fusion,” Information Fusion, 2017.

[7] A. Bhavan et al., “Cross-attention fusion audio-visual model for discrete emotion recognition on RAVDESS and CREMA-D,” Engineering Applications of Artificial Intelligence, 2023.

[8] T. Mittal et al., “M3ER: Multiplicative Multimodal Emotion Recognition Using Facial, Textual, and Speech Cues,” arXiv preprint, 2019.

[9] X. Wang & X. Zhao, ibid., emphasis on end-to-end deep methods using ResNet, CNN, and LSTM.

[10] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 527–536. doi: 10.18653/v1/P19-1050.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90

[12] S. Zagoruyko and N. Komodakis, “Wide residual networks,” in Proc. British Machine Vision Conf. (BMVC), York, UK, 2016, pp. 87.1–87.12. doi: 10.5244/C.30.87

[13] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.

[14] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, Aug. 2017, doi: 10.1016/j.neunet.2017.02.013.

[15] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017, doi: 10.1016/j.inffus.2017.02.003.

[16] H. Wang, A. Datta, and M. T. Ghassemi, "Multimodal emotion recognition using deep learning," IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 650–662, Apr.–Jun. 2022, doi: 10.1109/TAFFC.2020.2976576.

[17] Z. Zhang, J. Han, and L. Wang, "Multimodal learning for emotion recognition: A survey," Pattern Recognition Letters, vol. 145, pp. 123–132, Mar. 2021, doi: 10.1016/j.patrec.2021.01.016.

[18] Y. Wu, Q. Mi, and T. Gao, “A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions,” Biomimetics, vol. 10, no. 7, art. 418, Jul. 2025.

[19] D. Liu, Z. Wang, L. Wang, and L. Chen, “Multi-modal fusion emotion recognition method of speech expression based on deep learning,” Frontiers in Neurorobotics, vol. 15, art. 697634, Jul. 2021.

[20] P. Guo, Z. Chen, Y. Li, and H. Liu, “Audio-visual fusion network based on 3D-CNN and Conformer for multimodal emotion recognition,” in Lecture Notes in Computer Science, Jan. 2023.

[21] Y. Zhao, X. Cao, J. Lin, D. Yu, and X. Cao, “Multimodal affective states recognition based on multiscale CNNs and biologically inspired decision fusion model,” arXiv preprint, Nov. 2019.

[22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: from unimodal analysis to multimodal fusion,” Information Fusion, vol. 37, pp. 98–125, Sep. 2017.

[23] C. Woo Lee, K. Y. Song, J. Jeong, and W. Y. Choi, “Convolutional attention networks for multimodal emotion recognition from speech and text data,” arXiv preprint, May 2018.

[24] H. Wang et al., “Enhanced multimodal emotion recognition in healthcare analytics,” Elsevier, 2024.

[25] G. Caridakis et al., “Modeling naturalistic affective states via facial and vocal expressions recognition,” Proc. Second Int. Conf. Automatic Face and Gesture Recognition, Oct. 1996.

[26] C.-W. Wu and W.-B. Liang, “Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels,” IEEE Trans. Affective Computing, Jan. 2011.