iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78136

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Multimodal Emotion Classification from Audio and Video Using Hybrid Deep Learning Architectures on the CREMA-D Dataset

Sumedha Arya

Abstract: ***Emotion recognition is important for improving human computer interaction. In this study, we proposed two hybrid deep learning architectures are to recognize and classify six emotions such as anger, disgust, fear, happy, neutral, and sad using both audio and video data from the CREMA-D dataset.The first model uses a late fusion approach that combines ResNet-18 for extracting features from audio spectrograms, Vision Transformer (ViT-Tiny) for extracting features from video frames, and LSTM to learn the temporal patterns in video sequences. The second model replaces ResNet-18 and ViT with EfficientNet-B0 for both audio and video feature extraction, followed by LSTM for temporal learning and feature fusion.The results show that the EfficientNet-based model performs better, achieving an accuracy of 82%, while the ResNet18–ViT–LSTM model achieved 79% accuracy. The models performed very well in recognizing the happy emotion, but emotions like fear and sad were more difficult to classify.Overall, the results demonstrate that combining audio and visual information with temporal modeling can significantly improve emotion recognition performance.***

Keywords: ***Multimodal Emotion Recognition, CREMA-D Dataset, EfficientNet, ResNet-18, Vision Transformer, LSTM.***

I. INTRODUCTION

Emotionsplay an important role in human life. They influence how people think, behave, make decisions. This helps in understanding the world around them [12]. During normal conversations also, emotions help in deciding the tone of the discussion, the topics being talked about, and response of individuals to each other. Humans are naturally good at recognizing emotional signals from facial expressions, voice, and behavior. Therefore, based on these emotional cluses, they often adjust their actions.

According to authors [20], emotions can be explained using two main factors. These are called as valence and arousal. The valence describes whether an emotion is positive or negative, for example, it can be happiness and sadness, while the arousal represents intensity or strength of the emotion, ranging from low to high. In another study, authors [2] explained that the emotions help people quickly prepare for important social interactions. Some emotional responses are natural as they are inherited through evolution. On the other hand, some emotions are learned through experience and social interaction.

Based on the studies, done on emotions by Paul Ekman, six basic types were identified. These are happiness, sadness, anger, fear, surprise, and disgust [1]. Such emotions are commonly found across different cultures, society, and groups although the way people express them may vary.

With the rapid advancements in technology, human–machine interaction (HMI) is becoming more common. The advent of chatbots like ChatGPT, Gemini, Grok, Claude, and other virtual assistants, intelligent systems, there has been increase in the interaction of machines with humans in everyday life. For these systems to provide a better user experience, and response, it is important for them to understand the human emotions better [16]. Although not every machine requires to get smart in emotional understanding, but those systems who can recognize it well can adapt their behavior to the user's feelings and context. This can leadto more natural and effective interactions [17].Therefore, enabling machines to detect and understand human emotions has become acore area of research in fields such as computer vision, natural language processing and artificial intelligence.

II. LITERATURE REVIEW

Facial Emotion Recognition (FER) is one of the important tasks in computer vision. Previously, traditional machine learning algorithms such as Support Vector Machines (SVM) and logistic regression were used for emotion classification. However, with the increase in data, its pattern detection becomes challenging. Therefore, deep learning is used nowadays, replacing the traditional approaches by neural network–based models that can automatically learn complex features from data [11, 22].

©IJRASET: All Rights are Reserved

Current FER methods are categorized into two main factors: temporal information and data modality. Temporal information is how much time-related information the model processes, while modality refers to the type of data used for emotion recognition. These factors play an important role in designing FER systems and influence their complexity and performance.

Based on temporal information, FER models can be divided into two categorical approaches as static and dynamic [11]. Static approaches analyze a single image at a time. These models often use additional information such as gender, age, or head pose to improve emotion recognition performance [27]. In comparison, dynamic approaches analyze sequences of images or videos, focussing on capturing changes in facial expressions over time. Several techniques have been proposed based on this approach, including temporal restricted Boltzmann machines [26] and 3D convolutional neural networks (3D-CNNs) [10].

To better capture temporal dependencies, sequential models as recurrent neural network architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are widely used in FER tasks [24, 25]. These modelsprocess data sequentially and learn temporal patterns from video frames. Improved versions such as Bidirectional LSTM (BiLSTM) and Bidirectional GRU (BiGRU) process sequences in both directions, that means forward and backward to capture more contextual information. For example, BiLSTM has been combined with convolutional networks to detect pain intensity from facial expressions [4]. Similarly, some studies use separate LSTM networks for different modalities such as audio and visual data to perform emotion recognition from videos [14]. More recently, researchers proposed FER systems that combine convolutional layers, batch normalization, and ReLU activation to improve performance in real-world conditions [8].

Another way to categorize FER methods is based on modalities. It can be unimodal or multimodal in nature. In unimodal approaches, the model uses only one type of data, usually visual facial images [21]. Even when audio is used, it may be converted into spectrogram images for feature extraction, although the modality still remains same [9]. In contrast, multimodal approaches combine information from multiple sources such as facial expressions, speech, and text. In multimodal systems, different fusion techniques are used. These are sensor-level fusion, where raw data from different sensors is combined [23], feature-level fusion, where extracted features from different modalities are combined within the network [7], score-level fusion, where prediction scores from different models are combined [3], and decision-level fusion, where final decisions from multiple models are combined using voting mechanisms [13].

In FER, despite of progress in research, one major challenge still exists, that is the lack of realistic datasets. Many existing datasets are collected in controlled laboratory environments, such as IEMOCAP Dataset [5] and RECOLA Dataset [19]. While these datasets provide useful training data, but they may not fully satisfy real-world situations. On the other hand, “in-the-wild” datasets such as MELD Dataset [18] and MUStARD Dataset [6] are often collected from movies or television shows. Although they appear more realistic, however professional actors were involved in data collection along with controlled recording conditions, which can limit the generalization ability of models in real-world environments [15].

Overall, recent FER research mainly focusses on deep learningbased multimodal systems that can process temporal information from videos. These systems often use techniques such as transfer learning, convolutional neural networks, and recurrent neural networks like LSTM, BiLSTM to improve emotion recognition performance. In many cases, multimodal datasets in form of audio and visualfeatures were extracted from videos and are combined at the feature level to build more robust FER models.

III. RESEARCH METHODOLOGY

This section describes the research methodology used for classification of emotional multimodal data comprising of audio and video details of CREMA – D dataset. The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) with a primary focus on audiovisual emotion cues (facial expressions and vocal prosody). The various steps used in methodology are dataset acquisition, preprocessing, exploratory data analysis (EDA), and statistical visualization. The implementation is carried out using Python 3 in a Kaggle notebook environment with GPU acceleration enabled.

A. Dataset Description

The CREMA-D datasetserves as the primary data source. It is a multimodal (audio-visual) corpus specifically designed for emotion recognition research, containing 7,442 short video clips recorded from 91 professional actors (48 male, 43 female) aged 20–74 years, representing diverse ethnic backgrounds (African American, Asian, Caucasian, Hispanic).

Each actor performed 12 fixed English sentences, expressed in six categorical emotions:

- Anger (ANG)
- Disgust (DIS)

- Fear (FEA)
- Happy (HAP)
- Neutral (NEU)
- Sad (SAD)

Emotions were enacted at varying **intensity levels**:

- Low (LO)
- Medium (MD)
- High (HI)
- Unspecified (XX)

Video files follow a structured naming convention: ActorID_Statement_Emotion_Intensity_Gender.mp4 (e.g., 1001_DFA_ANG_XX_01.mp4 → Actor 1001, sentence DFA, Anger, unspecified intensity, male).

The dataset is publicly available under the Open Database License and was accessed via Kaggle at: /kaggle/input/datasets/alenken/multimodal-emotion-recognition-ravdess/crema-d

B. Data Acquisition

The dataset was loaded by traversing the directory structure using os.walk() to identify all .mp4 files. Each filename was parsed to extract the following metadata:

- video_path: Full path to the video file
- actor_id: Unique actor identifier (1001–1091)
- statement: Sentence code (e.g., IEO, ITS, MTI, DFA, etc.)
- emotion: Emotion label (ANG, DIS, FEA, HAP, NEU, SAD)
- intensity: Intensity level (LO, MD, HI, XX)
- gender: Derived from the last code (01 = male, 02 = female)

The extracted information was stored in a Pandas DataFrame (df) with 7,442 rows and 6 columns. No additional preprocessing steps were performed in this phase, as the current work focuses on dataset understanding and distribution analysis.

C. Exploratory Data Analysis (EDA)

Exploratory analysis was conducted to understand the distribution and balance of key categorical variables:

- Statement (sentence code)
- Emotion
- Intensity
- Gender

For each variable, the following analyses were performed:

- Unique values — to identify all possible categories.
- Value counts — to compute frequency of each category.
- Bar plot — to visualize absolute counts.
- Pie chart — to visualize percentage distribution.

These visualizations were generated using Matplotlib.

D. Implementation Environment

- Programming Language: Python 3.12
- Core Libraries:
  - numpy, pandas — data manipulation
  - os — file system traversal
  - matplotlib.pyplot — statistical visualization
- Execution Platform: Kaggle Notebook (NVIDIA Tesla T4 GPU accelerator enabled)
- Reproducibility: All paths are relative to the Kaggle input directory; random seeds were not set as no stochastic operations (e.g., model training) were performed in this stage.

- Data Preprocessing: For the visual modality, a fixed sequence of eight frames is selected at equal intervals from video using OpenCV. Each frame is then processed through several transformations to clean images for modeling. For the audio modality, the processing of corresponding audio signal is performed. Theyare first resampled to a standard sampling rate of 16 kHz to ensure consistency across all samples. A Mel spectrogram is then generated with 64 Mel frequency bins and a Fast Fourier Transform (FFT) window size of 1024. The spectrogram is converted to the decibel scale to better represent sound intensity variations. Finally, the spectrogram is resized to a fixed dimension of 64 × 128 so that it can be used as an input feature representation for the deep learning model.

E. Proposed Hybrid Model 1 – Resnet18, ViT and LSTM

The proposed multimodal architecture uses a late fusion strategy to combine information from both modalities. For video feature extraction, a Vision Transformer Tiny (ViT-Tiny) model is used to obtain frame-level features from each video frame. Since emotions often involve temporal dynamics, the extracted frame features are passed to a Long Short-Term Memory (LSTM) network, which models sequential relationships between frames and produces a compact representation of the video sequence. For the audio branch, a ResNet-18 convolutional neural network is used to process the Mel spectrogram and extract high-level audio features representing emotional patterns in speech.

After extracting features from both modalities, the outputs from the video and audio networks are concatenated to form a combined multimodal representation. A dropout layer with a rate of 0.5 is applied to reduce overfitting and improve generalization. The fused feature vector is then passed through a fully connected classification layer that predicts the final emotion category.

The model is trained using the AdamW optimizer with a learning rate of $5 \times 10^{-5}$and a weight decay of 1×10^{-2}. Cross-entropy loss is used as the objective function for multi-class emotion classification. The training process is performed with a batch size of eight for up to ten epochs. To further prevent overfitting and reduce unnecessary training time, an early stopping mechanism is implemented. If the validation loss does not improve for three consecutive epochs, the training process is automatically terminated.

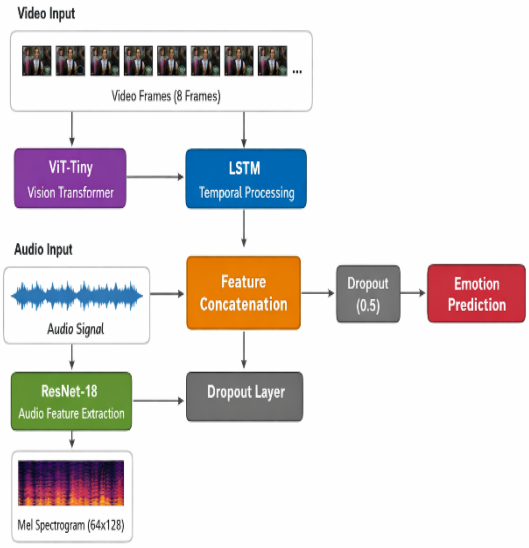


Fig. 1: Proposed Model Architecture (Resnet18+ViT+LSTM)

F. Proposed Hybrid Model 2 – EfficientNet and LSTM

The proposed multimodal architecture also utilizes a late fusion strategy to integrate information from both video and audio modalities. For video feature extraction, a pretrained EfficientNet-B0 convolutional neural network is used to extract spatial features from individual video frames. Since emotional expressions evolve over time, the extracted frame-level features are passed to a Long Short-Term Memory (LSTM) network. For the audio branch, the spectrogram-based speech signal is processed using another EfficientNet-B0 model. This network extracts high-level acoustic features that capture emotional characteristics.After extracting features from both modalities, the rest of the architecture is same as we used in proposed hybrid model 1 with Resnet18, ViT and LSTM.

©IJRASET: All Rights are Reserved

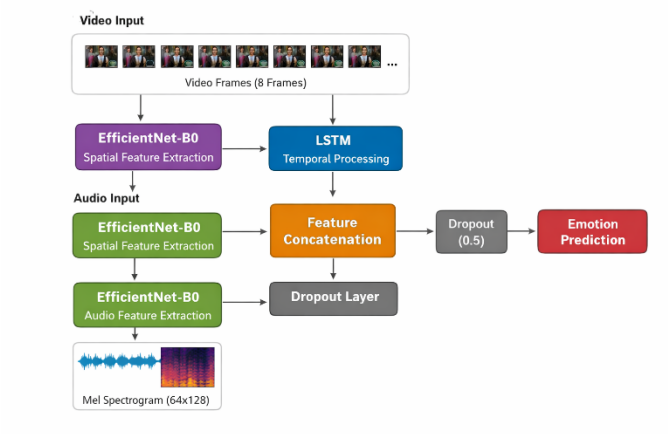


Fig. 2: Proposed Model Architecture (EfficientnetB0+LSTM)

IV. RESULTS ANALYSIS

The experimental results show that both proposed multimodal architectures are capable of recognizing human emotions from audio and video data. The Proposed Model 1, which combines ResNet18, Vision Transformer (ViT), and LSTM, achieved an overall accuracy of 79% on the test dataset. During training, the model showed steady improvement as the training loss decreased from 1.6131 to 0.4183, while the validation loss reduced from 1.3236 to 0.5887, indicating effective learning of emotional features. The model performed best in detecting the happy emotion with an F1-score of 0.94, followed by angry and disgust. However, lower performance was observed for emotions such as fear and sad, which are generally more complex and difficult data patterns to distinguish.

The Proposed Model 2, which uses an EfficientNet-based multimodal architecture with LSTM, achieved better performance with an overall accuracy of 82%. The training loss decreased significantly from 1.1102 to 0.0607, while the validation loss stabilized around 0.6125, indicating strong feature learning capability. Compared to Proposed Model 1, this model showed improved classification for several emotion classes, particularly fear, neutral, and sad. The happy emotion again achieved the highest performance with an F1-score of 0.96. Overall, the EfficientNet-based architecture demonstrated better generalization and improved multimodal feature extraction.

Table 1: Performance Comparison of Multimodal Models

Model	Architecture	Accuracy	Precision (Macro)	Recall (Macro)	F1-score (Macro)
Hybrid Model 1	ResNet18 + ViT + LSTM	0.79	0.79	0.79	0.79
Hybrid Model 2	EfficientNet + LSTM	0.82	0.82	0.82	0.82

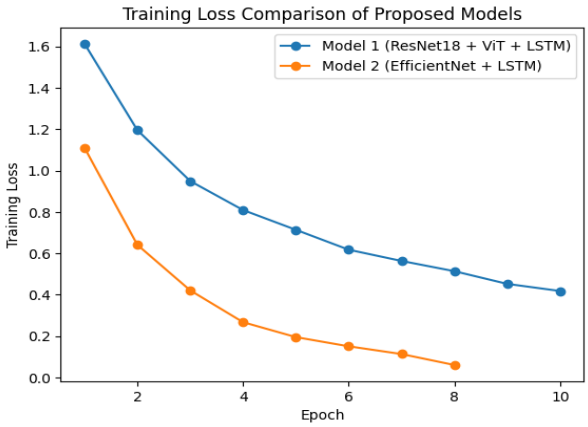


Fig. 3: Training Loss Comparison of Proposed Models

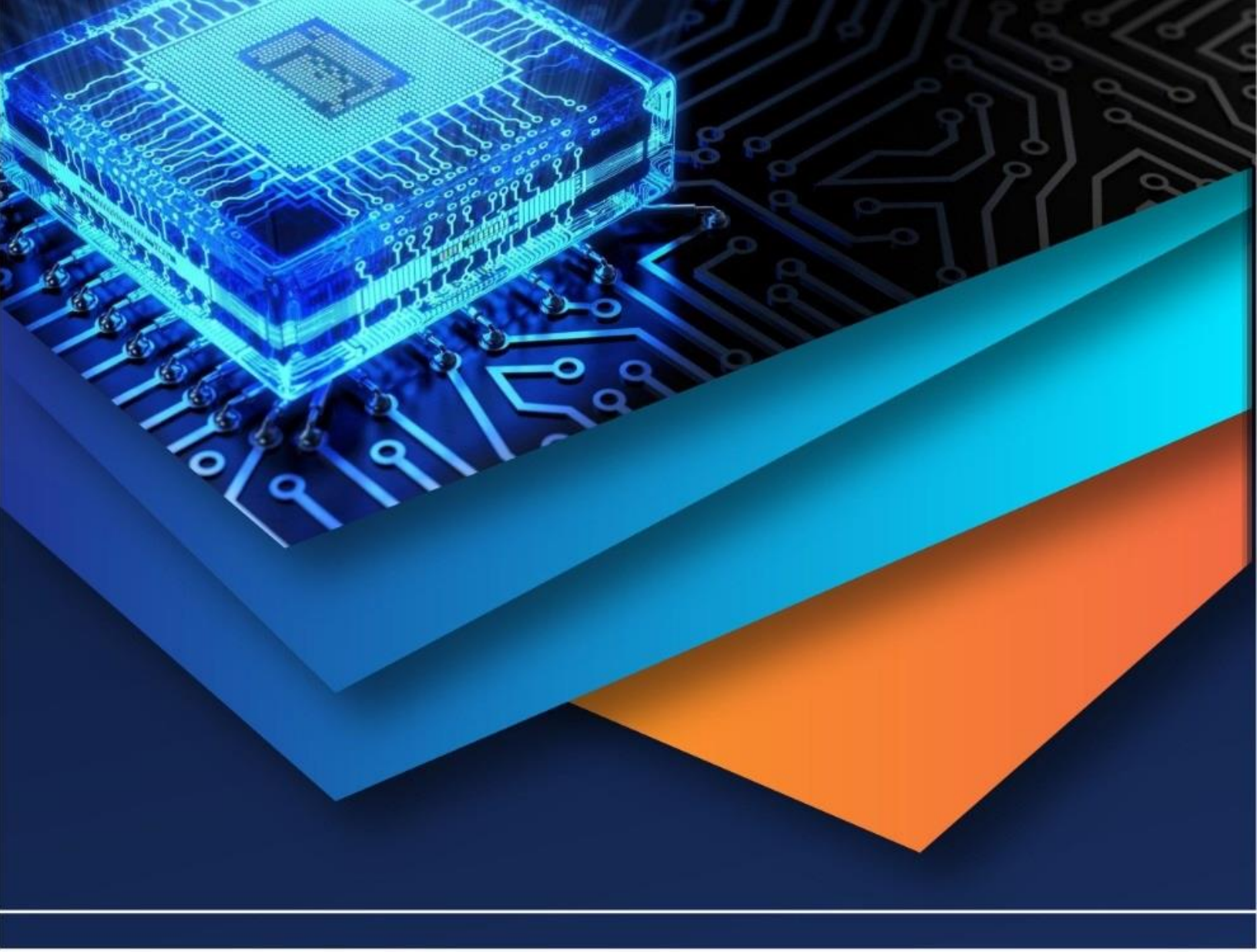
V. CONCLUSION

This study proposed two hybrid multimodal deep learning architectures for multi-class emotion recognition using audiovisual data. The first model integrates ResNet18, Vision Transformer, and LSTM, while the second model uses EfficientNet with LSTM for feature extraction and temporal modeling. Experimental results show that the EfficientNet–LSTM model outperforms the ResNet18–ViT–LSTM model, achieving a higher accuracy of 82% compared to 79%. Future work may focus on incorporating attention mechanisms, transformer-based temporal modeling, and larger multimodal datasets to further enhance the performance of emotion recognition systems.

REFERENCES

[1] P. Ekman, “An argument for basic emotions,” Cognition & Emotion, vol. 6, no. 3–4, pp. 169–200, 1992.
[2] P. Ekman et al., “Basic emotions,” in Handbook of Cognition and Emotion, pp. 45–60, 1999.
[3] K. Aizi and M. Ouslim, “Score level fusion in multi-biometric identification based on zones of interest,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 1, pp. 1498–1509, 2022.
[4] G. Bargshady, X. Zhou, R. C. Deo et al., “Enhanced deep learning algorithm development to detect pain intensity from facial expression images,” Expert Systems with Applications, vol. 149, p. 113305, 2020.
[5] C. Busso, M. Bulut, C. C. Lee et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.
[6] S. Castro, D. Hazarika, V. Pérez-Rosas et al., “Towards multimodal sarcasm detection (an obviously perfect paper),” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.
[7] H. Fan, X. Zhang, Y. Xu et al., “Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals,” Information Fusion, vol. 104, p. 102161, 2024.
[8] D. Freire-Obregón, D. Hernández-Sosa, O. J. Santana et al., “Towards facial expression robustness in multi-scale wild environments,” in International Conference on Image Analysis and Processing, 2023.
[9] J. Y. Kim and S. H. Lee, “CoordViT: A novel method to improve vision transformer-based speech emotion recognition using coordinate information concatenate,” in 2023 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1–4, 2023.
[10] S. Kumawat, M. Verma, and S. Raman, “LBVCNN: Local binary volume convolutional neural network for facial expression recognition from image sequences,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 207–216, 2019.
[11] S. Li and W. Deng, “Deep facial expression recognition: A survey,” IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, 2022.
[12] C. Lisetti, “Affective computing,” Pattern Analysis and Applications, vol. 1, pp. 71–73, 1998.
[13] S. Liu and R. He, “Decision-level fusion detection method of hydrogen leakage in hydrogen supply system of fuel cell truck,” Fuel, vol. 367, p. 131455, 2024.
[14] A. I. Middya, B. Nag, and S. Roy, “Deep learning based multimodal emotion recognition using model-level fusion,” 2022.
[15] B. Pan, K. Hirota, Z. Jia et al., “A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods,” Neurocomputing, vol. 561, p. 126866, 2023.
[16] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.
[17] R. W. Picard, “Toward computers that recognize and respond to user emotion,” IBM Systems Journal, vol. 39, no. 3–4, pp. 705–719, 2000.
[18] S. Poria, D. Hazarika, N. Majumder et al., “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 527–536, 2019.
[19] F. Ringeval, A. Sonderegger, J. S. Sauer et al., “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions,” in IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–8, 2013.
[20] J. A. Russell, “Pancultural aspects of the human conceptual organization of emotions,” Journal of Personality and Social Psychology, vol. 45, no. 6, p. 1281, 1983.
[21] E. Ryumina, D. Dresvyanskiy, and A. Karpov, “In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study,” Neurocomputing, vol. 514, pp. 435–450, 2022.
[22] M. Sajjad, F. U. M. Ullah, M. Ullah et al., “A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines,” Alexandria Engineering Journal, vol. 68, pp. 817–840, 2023.
[23] M. Singh, R. Singh, and A. Ross, “A comprehensive overview of biometric fusion,” Information Fusion, vol. 52, pp. 187–205, 2019.
[24] V. Vielzeuf, S. Pateux, and F. Jurie, “Temporal multimodal fusion for video emotion classification in the wild,” in Proceedings of the 19th ACM International Conference on Multimodal Interaction, New York, NY, USA, pp. 569–576, 2017.
[25] M. T. Vu, M. Beurton-Aimar, and S. Marchand, “Multitask multi-database emotion recognition,” in IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3630–3637, 2021.
[26] S. Wang, Z. Zheng, S. Yin et al., “A novel dynamic model capturing spatial and temporal patterns for facial expression analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 9, pp. 2082–2095, 2020.
[27] Z. Zhang, P. Luo, C. C. Loy et al., “From facial expression recognition to interpersonal relation prediction,” 2018.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)