ISSN: 2581-3404 (Online) **IF: 5.86 (SJIF)**
IJIRTM, Volume-10, Special Issue-2, April-2026

Real-Time Emotion Recognition Using Multimodal Deep Learning

***Mridul*[1], *Amit*[2], Monty[3], *Rajni*[4]**
Research Scholar, Department of Computer Science & Engineering (AI & DS), Panipat Institute of Engineering and Technology, Panipat, India [1,2,3,4]
mridulrohilla012@gmail.com[1],*mail4amit04@gmail.com*[2],*montyysaini@gmail.com*[3]

Abstract. *The use of emotion recognition in human computer interaction, mental health, intelligent agents and adaptiveAI systems is in a number of ways.We propose a real-time multi modal emotion recognition system which use facial expressions, speech signals, body language and textual inputs using deep learning in this paper. It is based on a mixture of CNN and LSTM networks, transformer-based NLP models, and posees timation. It uses a late fusion approach to make cross-modal predictions. The outcomes of the experiment prove that the suggested system achieves the average performance of 91.4% and the regular detectability under various environmental conditions. This is a Python-based system designed to be deployed in real-time.*

Keywords: Emotion Recognition, Multimodal Learning, Computer Vision, NLP, Speech Processing, Deep Learning, Python.

Introduction

Recognition of Emotion has emerged to be a critical component of Intelligent systems. Mental health monitoring to education, smart assistants and surveillance systems. State of the art approaches incorporate primitive modalities, e.g. speech or facial expression recognition algorithms, which in unfavorable scenarios fail because of noise and obscuration by the surroundings.

Human emotion is multiplex and expressed in terms of facial expression, voice, gesture and words. Having such a multimodal approach provides better generalization and robustness. New trends in deep learning have allowed integrating various modalities into one system.

Our proposed multimodal emotion recognition system is real time, in Python implementations. I Feature Aapplies the modalities of input(i.e.video, audio, body posture and text) concurrently in the system and provides a combined emotional state prediction.

2. Related Work

Previous methods need edhand-constructed features like Histogram of Oriented Gradients (HOG) for face and MFCC features for speech analysis. But they were not strong in the field.

The deep learning algorithms like CNNs, RNNs, and transformers have greatly enhanced performance. Multimodal methods have shown to be more accurate but can be plagued by the complexity of integration.

Table 1: Summary of Related Work

Sno.	Year	Papers	Focus	KeyFinding	Limitation
[1]	2016	K rizhevskyetal.–Deep CNNs	Facial Emotion Recognition	CNNs significantly improve image -based emotion detection accuracy	Requires large labeled datasets and high compute
[2]	2018	Tri georgisetal.–End-to- End Speech Emotion Recognition	Speech Processing	Captures temporal emotional patterns using deep models	Sensitive to background noise
[3]	2020	P oriaetal.–Multimodal Emotion Recognition	Multimodal Learning	Fusion of audio-visual data improves accuracy	Limited modalities(no body/text)
[4]	2021	D evlinetal.–BERT Model	NLP/Text Emotion	Context-aware sentiment understanding	High computational cost
[5]	2019	Caoetal.– Open Pose-Based Emotion Recognition Z hangetal.–Multimodal Fusion Networks	Body Language Analysis	Body posture contributes to emotion detection	Less a ccuratein crowded scenes
[6]	2022	Lietal.–Real-Time Emotion Systems	Deep Fusion Techniques	Late fusion improves robustness across modalities	Complex model architecture
[7]	2023		Real-Time AI Systems	Achieves low latency with optimized pipelines	Trade-off between Speed and accuracy

3. Methodology

A. Dataset

To guarantee diversity and robustness-FER 2013(Facial), RAVDESS(Speech), Open Posedatasets(Body Language), Twitter datasets (Text) multiple datasets were combined.

Table 2: Data set Composition

Incident Category	Train	Validation	Test
Face	8000	2000	1500
Speech	6000	15000	1000
Body	5000	12000	800
Text	10000	25000	1500

B. Face Emotion Recognition

It has a CNN-based architecture, which comprises of convolutional, pooling and fully connected layers.Spatial features, including muscle facial motions, are extracted in the model.

C. Speech Emotion Recognition

AudiosignalsareconvertedintoMel-spectrogramsandprocessedusingLSTMnetworkstocapturetemporaldependencies.

D. Body Language Analysis

Pose estimation using Open Pose extracts key points such as joints and posture. These features are passed into a classification network.

E. Text Emotion Detection

Text input is processed using tokenization, embeddings, and transformer-based models such as BERT for sentiment analysis.

F. Multimodal Fusion

Late fusion is used where predictions from each modal it yare combined using weighted averaging.

4. System Architecture & Implementation

The system is divided into five major layers: Input, Preprocessing, Model, Fusion, and Output.

Table 3: System Layers and Responsibilities

Layer	Module	Responsibility
Input	Web UI, Camera, Microphone	Captures video, audio, and text input from user
AI Detection	CNN(Face), LSTM (Speech), BERT(Text),	Detects emotion features from each modality

Processing	Open Pose (Body) Fusion Engine, Confidence Module	Combines outputs and computes final emotion
Data	Database(SQ Lite)	Stores emotion logs and system records
Output	UI Dashboard, Report Generator	Displays emotion result and generates reports

5. Results and Discussion

Performance Metrics

All emotion categories, as discussed in Table 4 and illustrated here, show strong performance of the system. In general emotions like surprise end happy show higher precision and recall as they are depicted in different facial and vocal patterns hence have higher F1 scores. On the other hand, emotions such as fear and sad do not perform as well due to finely grained and shared features across modalities. The overall means performance of 91.4% accuracy shows that compared to uni modal systems, the multimodal one not ably increases the reliability of detection. By integrating face, speech, body language and text, the system can fill in when one modality is missing or ambiguous.

Table 4: Per Category Detection Performance

Incident Category	Precision	Recall	F1	AP@0.5
Happy	0.93	0.91	0.92	92.4%
Sad	0.90	0.88	0.89	89.7%
Angry	0.91	0.89	0.90	90.3%
Surprise	0.94	0.92	0.93	93.1%
Fear	0.88	0.86	0.87	87.9%
Neutral	0.92	0.90	0.91	91.2%
Mean(overall)	0.91	0.89	0.90	91.4%

Comparative Analysis
The presented system is compared with five alternatives presented in Table 5. The proposed system is more accurate compared to the rest alternatives as indicated in the table.

Table 5: Comparison of Detection Approaches

Method	Accuracy	Inference(ms)	Location
Manual Observation	Variable	Minutes 25	NoNoNo
Facial Emotion(CNN only)	90.2%		
Speech Emotion(LSTM only)	88.5%	30	
Audio-Visual(Face + Speech)	89.7%	35	Partial
Transformer-based NLP(Text only)	92.1%	40	No
Proposed(Multimodal System)	91.4%	21	Yes

Latency Analysis
The system achieves an average inference time of 2.1 seconds, suitable for real-time applications.

6. Conclusion and Future Scope
This paper presents a holistic multimodal emotion recognition system, whose basis is a face, speech, body language and text. The accuracy and robustness of the system compared with the traditional methods are better. It is used in the future to put it on the edge and combine it with the wearable sensors and real-time streaming applications.

Acknowledgment
The authors would like to than kimmensely the irfaculty mentors and institution to go with them through many of the stages of this working research. Other resources that have inspired this research are the open-source community, and numerous researchers in the field of computer vision, smart city systems, and AI.

References:
1.1
[1] P.EkmanandW.V.Friesen,FacialActionCodingSystem,ConsultingPsychologistsPress,1978.
[2] Y.LeCun,Y.BengioandG.Hinton,“Deeplearning,”Nature,vol.521,no.7553,pp.436–444,2015.
[3] I.Goodfellow,Y.BengioandA.Courville,DeepLearning,MITPress,2016.
[4] A.Krizhevsky,I.SutskeverandG.Hinton,“ImageNetClassificationwithDeepCNNs,”NIPS,2012.

[5] A.Graves,A.MohamedandG.Hinton,“Speechrecognitionwithdeeprecurrentneuralnetworks,”ICASSP,2013.
[6] K.SimonyanandA.Zisserman,“VeryDeepConvolutionalNetworks,”ICLR,2015.
[7] J.Devlin,M.Chang,K.LeeandK.Toutanova,“BERT:Pre-training of Deep Bi directional Transformers,” NAACL, 2019.
[8] A.Vaswanietal.,“AttentionIsAllYouNeed,”NeurIPS,2017.
[9] S.HochreiterandJ.Schmidhuber,“LongShort-TermMemory,”NeuralComputation,1997.
[10] Z.Zhang,P.LuoandX.Wang,“MultimodalEmotionRecognition:ASurvey,”IEEETransactionsonAffectiveComputing, 2020.
[11] S.Poria,E.Cambria,R.BajpaiandA.Hussain,“AReviewofAffectiveComputing,”InformationFusion, 2017.
[12] K.K.Kimetal.,“MultimodalEmotionRecognitionUsingDeepNeuralNetworks,”IEEEAccess,2018.
[13] A.Mollahosseini,D.ChanandM.Mahoor,“Goingdeeperinfacialexpressionrecognition,”WACV,2016.
[14] S.LiandW.Deng,“DeepFacialExpressionRecognition:ASurvey,”IEEETransactionsonAffectiveComputing,2020.
[15] G.Baltrušaitis,C.AhujaandL.Morency,“MultimodalMachineLearning:ASurvey,”IEEETPAMI,2019.
[16] Z.Caoetal.,“OpenPose:RealtimeMulti-Person2DPoseEstimation,”IEEETPAMI,2019.
[17] S.R.LivingstoneandF.A.Russo,“TheRyersonAudio-Visual Database of Emotional Speech and Song (RAVDESS),”
PLOSONE,2018.
[18] I.Goodfellowetal.,“ChallengesinRepresentationLearning:FER2013Dataset,”ICMLWorkshop,2013.
[19] E.Cambria,B.Schuller,Y.XiaandC.Havasi,“NewAvenuesinOpinionMiningandSentimentAnalysis,”IEEE Intelligent Systems, 2013.
[20] M.Wöllmeretal.,“Context- sensitive multimodal emotion recognition,” IEEE Transactionson Affective Computing,2013.