

# Multimodal perception for enhancing human computer interaction through real-world affect recognition

Karishma Raut<sup>1</sup>, Sujata Kulkarni<sup>1</sup>, Ashwini Sawant<sup>2</sup>

<sup>1</sup>Department of Electronics and Telecommunication Engineering, Sardar Patel Institute of Technology, Mumbai, India

<sup>2</sup>Department of Electronics and Telecommunication Engineering, Vivekanand Education Society's Institute of Technology, Mumbai, India

## Article Info

### Article history:

Received Jul 15, 2024

Revised Oct 17, 2024

Accepted Oct 30, 2024

### Keywords:

Affect recognition

Audio-visual data

Deep neural network

Multimodal

Preprocessing

## ABSTRACT

Human-Computer Interaction can benefit from real-world affect recognition in applications like healthcare and assistive robotics. Human express emotions through various modalities, with audio-visual being the most significant. Using a unimodal approach, such as only speech or visual, is challenging in natural, dynamic environments. The proposed methodology integrated a pretrained model with a convolution neural network (CNN) to provide a robust initialization point and address the limited availability of facial expression data. The multimodal framework enhances discriminative power by combining visual scores with speech. This work addresses the challenges at each stage of the real-world affect recognition framework, including data preprocessing, feature extraction, feature fusion, and final classification. A 1D-CNN is employed for training on spectral and prosodic audio features, while deep visual features are processed using a 2D-CNN. The proposed system's performance was evaluated on the extended Cohn-Kanade (CK+), acted-facial-expressions in-the-wild (AFEW), and real-world-affective-face-database (RAF) datasets, which are commonly used in face recognition research. Experimental results indicate that 2% to 5% of visual data from natural settings were undetected, and the inclusion of the audio modality improved performance by providing relevant and supplementary information.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Karishma Raut

Department of Electronics and Telecommunication Engineering, Sardar Patel Institute of Technology

Mumbai, India

Email: karishma.raut@spit.ac.in

## 1. INTRODUCTION

Affect recognition with artificial intelligence and machine learning techniques has shown tremendous applications in assistive robotics, healthcare, and security [1]–[3]. The analysis of human behavior using artificial intelligence and machine learning techniques can benefit society in a variety of ways. Children with autism may benefit from the development of human-computer interfaces (HCI) that can assist them in interpreting the facial expressions of others. An automated emotion recognition-based health assessment in a networked healthcare system provides telemedicine and a distant diagnostic or primary screening facility [2]. The humans use different modalities to express their emotions like, face, speech, body gestures. The audio and video signals non-intrusively explain how humans express themselves through facial reflections and speech in reaction to the occurrence of an emotion. In the field of affective computing, multimodal emotion recognition is a developing research area. It is difficult to investigate affect information from the dynamic natural environment of the real world. Because of occlusions, variations in head angle,

illumination conditions, and noisy environments, unimodal techniques are unable to deliver precise information. Real-world emotion recognition can be benefitted by multimodal learning that uses the complimentary data of various modalities to increase comprehension of emotions. The most popular non-invasive mode that can provide rich contextual information is audio-visual data. When both video and audio modalities are present, the human ability to perceive intended emotions is increased. In comparison to other input modalities, these two are more expressive and may be recorded without being obtrusive. As an example, visual information is critical for recognizing happiness, while audio information is essential for comprehending anger [4]. The way characteristics are taken from these two signals and fused together determines the recognition rate and robustness.

Early studies on affect recognition focused mostly on unimodal methods, such as speech and facial expression recognition on lab-controlled databases like extended Cohn-Kanade (CK+). These datasets are developed under regulated conditions, with maintained reverberation and noise level, typically with minimum speech content, illumination, and calibrated cameras [5]–[6]. Tools created based on this information generally perform poorly when used with behavioral recordings captured in the wild since such settings are so unlike from real-world applications. According to human psychology, People's emotional experiences are multimodal, especially audio-visual. The system may find it difficult to gather the information it needs from auditory and visual inputs in a natural setting to improve its capacity to categories emotions appropriately [7]–[9]. It is necessary to determine which databases have the most relevant information as well as how characteristics from visual and aural data are retrieved and combined. The Figure 1 shows sample data from different scenarios. Figure 1(a) indicates data from controlled environment and natural environment data is as shown in Figure 1(b) and Figure 1(c).



Figure 1. Sample data indicating scenarios from controlled and natural environment (a) lab-controlled datasets (CK+) [5] (b) real-world datasets (AffWild2) [8] and (c) acted-facial-expressions in-the-wild (AFEW) [7]

The primary objective of current research is to create frameworks or approaches that can operate reliably in both controlled and uncontrolled environments. One method for addressing video-based multimodal expression recognition issues in both the lab-controlled setting and the real-world is multiple feature fusion. Different features that can discriminate emotions from different modalities can be used to define new feature space. In video sequences, dynamic textures may be extracted using feature descriptors. It is possible to extract spectral and prosodic information from audio [10].

Many researchers have worked on combinations on different features and feature extraction, fusion techniques to improve affect recognition. The audio-visual feature selection with keyframes. The authors have learned to find effective metric through multimodality to obtain discriminative score. Further temporal aggregation of modalities in asynchronous manner is also evaluated [11]–[13].

The usual techniques for assessing the performance of classifiers include cross-validation and dividing the dataset into training and testing sets. These outcomes might be inflated due to high degree of similarity and shared corpora. Particularly in instances where there is no reproducibility, it is challenging to predict how a particular classifier would respond in many kinds of circumstances. As these models develop using lab-controlled data, it was not able to sustain in real-world environment. It was highlighted that cross corpus evaluation which means use of multiple databases for evaluation is required to develop precise models that can sustain in different environments [14].

To extract pertinent information from real-world data, the researchers are using a variety of methods. To extract both local characteristics from frames and global features from video, a hybrid neural network is developed. The temporal features are combined with two latent features retrieved from the networks. It recovers latent features including minute information from each frame [15]. The comparative analysis on two different types of datasets AFEW and CK+ clearly indicates the challenges for real-world corpora. Like this, models may indeed be created by combining several visual classifiers and utilizing their

own capabilities. The auditory and visual modalities are evaluated separately and then integrated at the decision level. Here, a variety of fusion techniques can be evaluated [16].

Researchers have put forward several techniques by considering various ways that people express their emotions [17]. Other researchers have also worked with various databases to determine whether the same methodology works on various datasets. A comprehensive literature review highlights the need to explore preprocessing real-world data, as it presents different conditions compared to lab-controlled data. This stage is crucial for effective feature extraction within the framework. Additionally, integrating multimodal information is essential for accurate affect recognition in real-world environments. The paper contributes as follows:

- The difficulties in preprocessing real-world data are emphasized through investigations and comparative studies using various detectors on uncontrolled, real-world visual data.
- The influence of this preprocessing on extraction of features and key feature selection is examined by analyzing the outcomes from a pre-trained convolution neural network (CNN) model applied to the CK+, RAF and AFEW datasets.
- The relative importance of another modality (audio) is highlighted through multimodal comparative analysis in real-world scenarios.

The paper is organized as follows: section 1 discusses introduction and related work. Section 2 describes methodology for real-world affect recognition framework and section 3 results and discussion. Finally, the conclusion is provided in Section 4.

## 2. METHOD

Real-world audio-visual affect recognition framework, first separates audio and video by maintaining its bit rate and pre-process it as depicted by Figure 2. Further, features are extracted from both the modalities and can be fused in multiple ways for classification. The human emotions are cultural-specific and can vary person to person. To analyze human emotions globally the realistic dataset with global participants, different situations, without language barrier will play significant role. Hence, Real, spontaneous expressions of affect have received greater attention in recent years but limited in number.

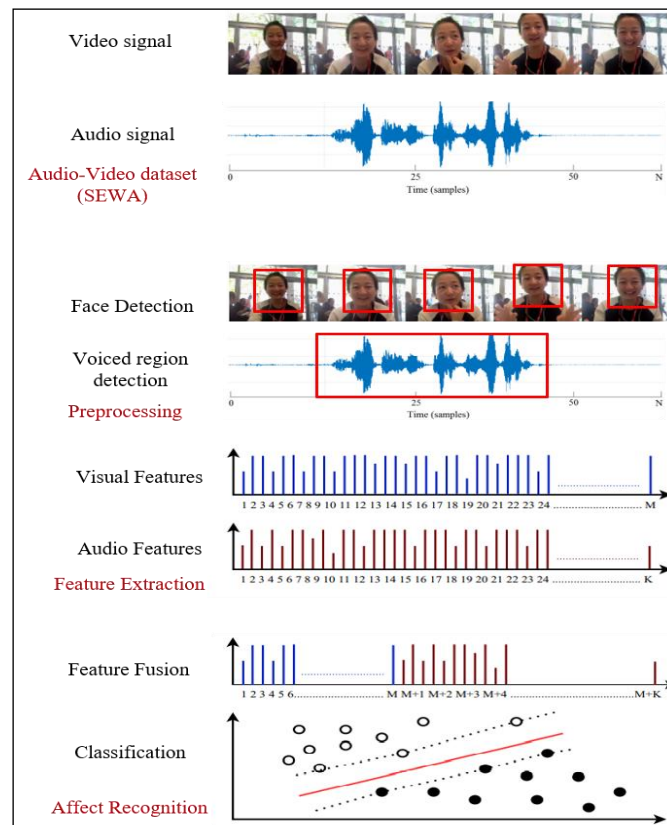


Figure 2. General work flow for audio-visual affect recognition

To capture the entire spectrum of expressions individuals, use in their daily lives, one method would be to combine data from several Internet sources. A few datasets with natural recordings have recently been accessible, including acted-facial-expressions in the wild (AFEW), affect-in-the-wild (AFF-WILD2), automatic-sentiment-analysis-in-the-wild (SEWA), and real-world-affective-face-database (RAF).

The dataset analysis of AFEW, Aff-Wild2, SEWA, and RAF by emotion class distribution, video duration, accessible video count, etc., is displayed in Figure 3. Two datasets can be discussed in light of the requirement for audio-visual data. AFEW contains more data about happiness, while Aff-Wild2 has the most frames of anger. Aff-Wild2 features lengthy sequences with frame-wise annotations. Comparing the distribution of emotion classes by video to AFEW, it is totally out of balance. It exhibits a non-uniform emotional distribution. Both datasets were acquired using the same way, which is from YouTube and movies. Hence, appropriate combination with respect to emotion categories can make a new combined dataset which may contribute more to recognize affect. The videos representing fear, disgust and surprise can be combined with AFEW data.



Figure 3. Real-world databases description

Preprocessing is a crucial stage where the key information is taken out of the video and audio modalities like face detection, alignment, voice or non-voice region detection [18]. The successful detectors like, OpenCV, single shot multibox detector (SSD), Multi-cascade convolution neural network (MTCNN) and RetinaFace are used in recent work [19]. It will have crucial impact on feature extraction if key areas are not detected properly. The audio-video signals are converted by feature extraction techniques into feature vectors for the classification model. Several methods are explored to find prominent features in terms of spatial, temporal, or spatiotemporal information from audio-visual data [20]. People in social situations express and perceive their emotional states across a range of distinct modalities. The robustness and accuracy of the model can be improved by multimodal fusion. It is crucial to research and evaluate how modalities should be combined. Fusion can be achieved either at the beginning of the model development process or later by integrating distinct models. The Table 1 explains preprocessing, feature extraction and fusion techniques.

Table 1. Preprocessing feature extraction and fusion techniques

Face detection and alignment		
Sr.No.	Method	Description
1.	OpenCV	It uses HAAR cascade classifier.
2.	SSD	The image is divided into grids, and each grid cell identifies objects in that area of the image.
3.	MTCNN	It is a deep convolutional neural network that can identify faces in coarse to fine spatial correlation with landmark locations.
4.	RetinaFace	It performs pixelwise face localization on various scales of faces.
Visual features		
Sr. No.	Method	Description
1.	Based on appearance	Changes in facial texture are concerned.
2.	Based on geometry	It refers to facial configurations in which a group of facial fiducial points helps identify the shape of the face.
3.	Based on dynamic texture-based	The dynamic motions and spatial appearance of a video series are simultaneously simulated. Ex: LBP-TOP, HOG-TOP
4.	Deep features	Relevant information is extracted using deep models.
Audio Features		
Sr. No.	Method	Description
1.	Prosodic features	They provide crucial indicators of the speaker's emotions and are impacted by the vocal folds. Ex: zero-crossing-rate (ZCR), short-time-energy (STE)
2.	Spectral features	They are governed by the vocal tract and communicate the frequency content of the speech stream. Ex: mel-frequency-cepstral-coefficient (MFCC) and linear-prediction-cepstral-coefficients (LPCC)
3.	Spectrogram	It shows the amplitude of the signal with time at different frequencies visually.
4.	Cochleagram	It is a Spectrogram variant that functions quite similarly to the human ear.
5.	Modulation spectral feature (MSF)	It is a sensory Spectro-temporal representation of the speech signal's long-term dynamics.
Fusion techniques		
Sr. No.	Method	Description
1.	Feature level or early stage	Near the raw sensor data, the features are extracted independently and combined. The model can learn how the modalities are correlated. Ex: principal component analysis, canonical correlation analysis, independent component analysis
2.	Decision level or later stage	Aggregates the results of independent recognition models optimally Ex: score-weighting, bayes-rules, max-fusion and average-fusion
3.	Intermediate	Data fusion at different stages of model training. Fusion layer or a shared representation layer.

## 2.1. Multimodal deep CNN framework

The multimodal framework shown in Figure 4 comprises visual data classification with deep CNN using FER pretrained model and audio data classification using CNN model trained by basic audio features. The results obtained from visual modality and audio modality can be analyzed separately to observe contribution of each and can be fused together for enhancing affect recognition.

The initial stage involves identifying and positioning the face since it has the most noticeable visual cues. The application of deep transfer learning through a suitable face alignment technique, closely resembles the pre-trained deep neural network's training environment for emotion recognition. The fact that there are more data available to train recognition models is one factor in this decision. Augmenting facial details for fine-grained classification is the fundamental idea behind both problems.

An open-source Python package called face emotion recognizer (FER) makes use of a CNN that was trained using the FER 2013 dataset [21]. The pretrained model trained on FER 2013 dataset provides rich features while training deep neural network with other dataset by overcoming scarcity of data. The 2D convolutional neural network is inspired from VGG-Face network. Each of the five convolutional layers with 256, 256, 128, 128, and 64 kernel respectively is succeeded by a max pooling layer. The final fully connected layer followed by SoftMax function for classification of seven emotions.

The performance of an algorithm created utilizing computer vision techniques may be impacted by several technical difficulties. As an illustration, different subjects may not express the same emotion the same way. Additionally, different points of view lead to disparate emotional representations. Furthermore, changes in illumination and occlusions may deceive the recognition mechanism. Important factors that could affect how well speech emotion recognition performs in the end include background noise and subject voice changes. The strategy of audio-visual impact recognition systems is to utilize the extracted information from one modality to enhance the recognition capabilities of the other modality by supplementing the missing information.

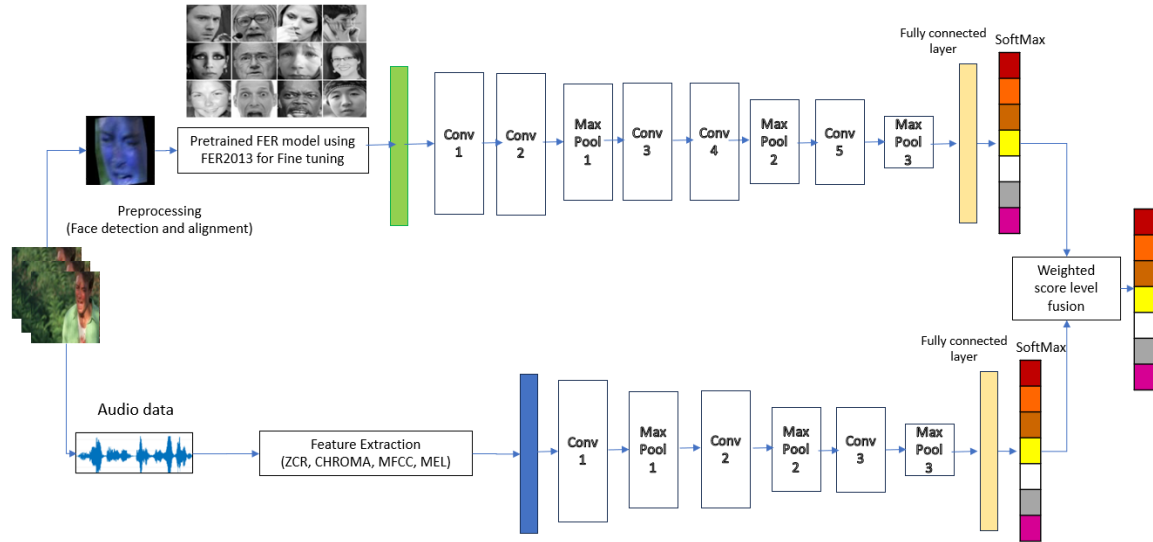


Figure 4. Multimodal deep CNN framework for real-world affect recognition

Based on several acoustic elements that correspond to each emotion, emotions may be defined. For instance, a quicker speaking tempo, greater energy level, and a higher pitch frequency can be used to convey anger. The key auditory characteristics for identifying emotions include pitch, speech intensity, speech duration, spectral-energy-distribution, mel-frequency-cepstral-coefficients (MFCCs), and average zero-crossing-rate (ZCR). These characteristics reflect the variations in the prosodic patterns of the voices of various speakers. The speaker's accent, speaking tempo, phrasing, speech pitch range, and intonation all determine the prosodic pattern. The identification of speakers can also benefit from the use of spectral properties that are taken from small speech data [22]–[23].

The 1D CNN model used for audio classification has three convolutional layers (256, 128, and 64 kernels), with a max pooling layer after each layer. The final fully connected layer followed by SoftMax function for classification of seven emotions. The weighted score level fusion is used for final classification with equal weights assign for both the modalities.

**2.2. Evaluation parameters for framework**

Only accuracy cannot be a measure of evaluation when data is unevenly distributed [23]. Hence appropriate evaluation is carried out using precision, recall and F1-score described in Table 2 [24]. The computation is as follows [25]:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{1}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{3}$$

$$F1 - score = \frac{(2*Recall*Precision)}{(Recall+Precision)} \tag{4}$$

Table 2. Evaluation parameters

Evaluation parameters	Description
Accuracy	It computes number of correct predictions across the entire dataset.
Precision	This refers to the percentage of actual positive observations inside the dataset that the model identified as positive.
Recall	percentage of positive observations relative to the total number of positive observations.
F1-score	It is the harmonic mean of the model's precision and recall, combining the model's accuracy and recall.

### 3. RESULTS AND DISCUSSION

#### 3.1. Experiment 1: preprocessing challenges

The impact of real-world diverse conditions is analyzed using different preprocessing techniques on visual data. The DeepFace python framework with VGG-FACE model is used to detect and align face from real-world dataset. The framework supports different face detection techniques under one roof.

The result for one complete video using OpenCV, SSD, MTCNN, RetinaFace detectors is shown in Figure 5. The impact of head rotation, change in illumination can be seen from shown results. The performance of RetinaFace and MTCNN is better than other detectors under dynamic conditions. The experimental analysis shows face detection and alignment is affected with change in illumination and head pose. The 5% data from AFEW dataset and 2% data from RAF dataset containing extreme head rotation, illumination are not detected by any detector. The low identification rate in real-world scenarios can be attributed to this, particularly. The comparative analysis of a lab-controlled dataset (CK+ with 981 images) and a real-world dataset (AFEW with 383 videos (5,838 diverse images) and RAF with 3,068 images) for affect recognition using pretrained CNN model (FER 2013) framework validates above discussion [20].

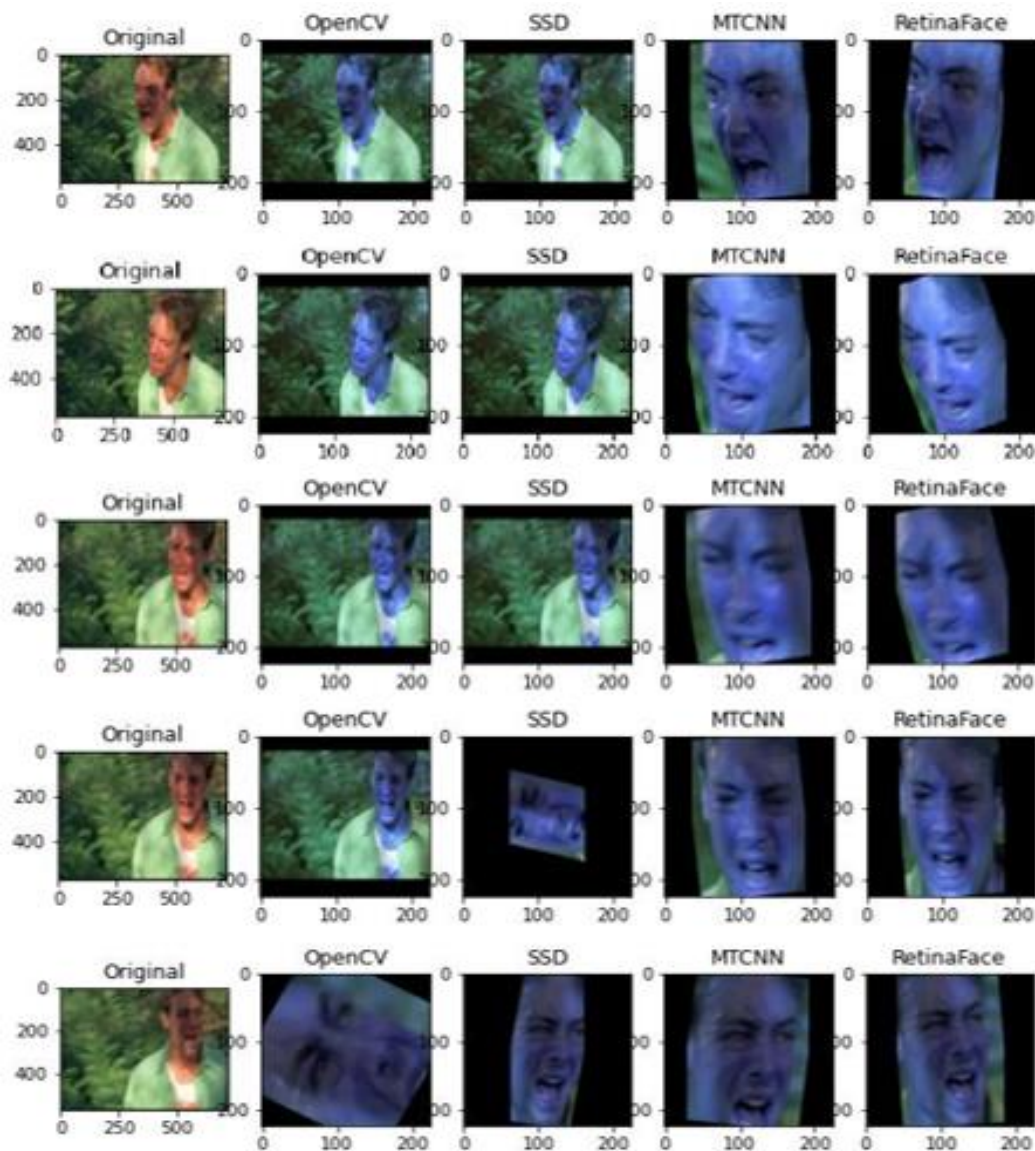


Figure 5. Impact on face detection and alignment with change in illumination and head pose for video sequence

**3.2. Experiment 2: comparative analysis of unimodal approach for affect recognition in lab-controlled and natural environment**

The preprocessing results from experiment 1 suggest use of cascaded MTCNN network to sustain in real-world dynamic conditions. The visual data for affect recognition using deep convolutional neural network with pre-trained CNN model is analyzed with a lab-controlled dataset CK+ and a representative wild dataset RAF and AFEW.

The experimental analysis shows following results represented in Table 3. The frontal posed samples of the CK+ dataset acquired in controlled environment is completely detected and having average recognition rate of 62.58%. The recognition rate of few emotions viz. happy, surprise, and sad is excellent as compared to other emotions due to clear facial actions and visibility. The model is trained by considering features such as the frown, the eyebrows and the eye widening, mouth opening. The interpreted results helped to understand the reasons behind some misclassification like, person wearing glasses being classified as angry assuming it is a frowning. Hence, angry, neutral have average performance and fear, disgust is almost misclassified. It can be noticed that the features learned are not sufficient to identify minute details of human emotions and hence, multiple features are required to enhance discrimination of emotions.

Table 3. Evaluation parameters of CK+, AFEW and RAF dataset

Dataset parameters	CK+	AFEW	RAF
Precision	0.571	0.25	0.393
Recall	0.598	0.27	0.479
F1-score	0.508	0.25	0.368
Accuracy	62.58%	31.07%	52.02%

The 5% data from AFEW dataset model was not detected by model and for rest it has 31.07% average recognition rate. The average recognition rate is observed for emotions happy, neutral, and sad. The fear and surprise are poorly recognized and disgust is completely misclassified. Around 2% data from RAF was not recognized and poor recognition rate can be observed for disgust and fear. The average recognition rate in terms of true positives for RAF is 52.02%. A facial expression may occasionally be confused with other expressions in the wild when numerous expressions coexist together. The extreme illumination conditions, head positions and occlusion are responsible for not detecting few data from dynamic dataset AFEW and RAF. Few samples of such images are depicted by Figure 6.

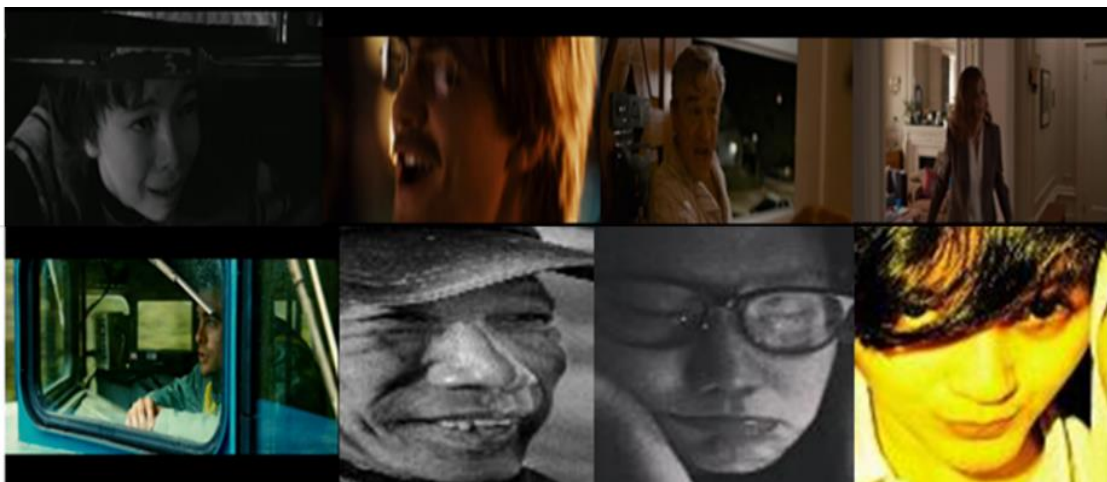


Figure 6. Sample images from RAF not recognized by any detector

The Figure 7 represents the sensitivity analysis of the described model. The model is trained and tested on AFEW, RAF and CK+. The experimental analysis demonstrates how challenging it may be to distinguish between an affective state in a real-world as compared to a controlled environment. It emphasizes the need for supplemental information to be provided via a variety of modalities.



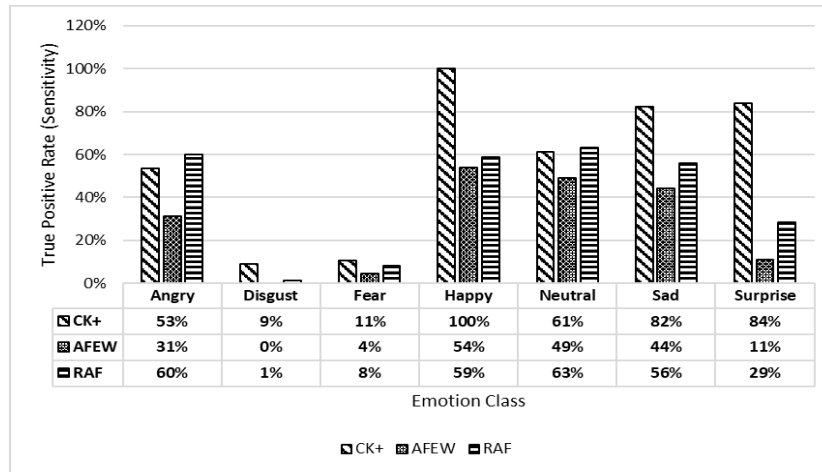


Figure 7. Sensitivity analysis of AFEW, RAF and CK+ with respect to emotion class

### 3.3. Experiment 3: performance analysis of multimodal approach in natural environment

A comparative analysis of unimodal and multimodal approaches is conducted in natural environment. The dataset used for training speech model is AFEW with 774 videos. The test AFEW (383 videos) dataset is same for audio and video model for classification. The results from the visual and audio modalities are analyzed separately to assess the contribution of each and then combined to improve affect recognition. The evaluation parameters are as shown in Table 4.

The Figure 8 shows sensitivity analysis of unimodal (audio and video) and multimodal (audio-video) approaches for emotion classes. Each modality is good at recognizing few emotions however combinedly all emotions can be efficiently recognized. It is observed that audio modality is excellent performance in recognizing anger whereas visual modality is good at recognizing happy. The emotion fear is recognized in better way when both modalities are present. The performance of surprise and disgust can be improved using more detailed features. As mentioned earlier, dataset contains less information about surprise and disgust and hence model is not trained in balanced way. The data augmentation can be applied to emotional data having less information. Also, different fusion techniques can be explored to enhance the affect recognition rate.

Table 4. Evaluation parameters for unimodal and multimodal framework with AFEW dataset

Modality parameters	Visual	Audio	Audio-video
Precision	0.25	0.24	0.31
Recall	0.27	0.30	0.29
F1-score	0.25	0.24	0.28
Accuracy	31.07%	33%	34.98%

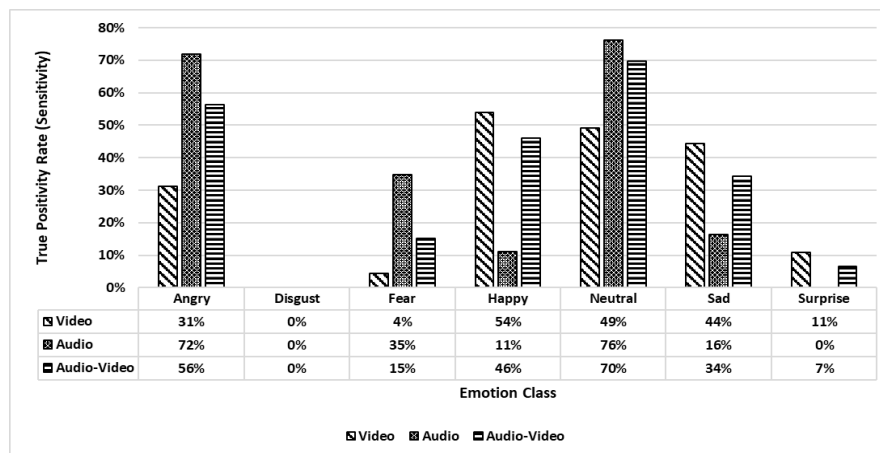


Figure 8. Sensitivity analysis of unimodal and multimodal approaches for emotion classes

The take away from above discussion is summarized as follows:

- The unimodal analysis is essential to understand contribution of each modality as well as problems encountered by the modality.
- The impact analysis of preprocessing is guiding for feature extraction and requirement of additional modality.
- The way data augmentation applied to enhance information can be changed and more weightage can be provided to classes with less data.

#### 4. CONCLUSION

The real-world situations present greater difficulties for face detection and alignment as compared to controlled situations. Due to occlusion and non-frontal head posture, the original facial expression could appear differently. It has severe impact on the misclassification of different emotions specifically anger, fear, surprise, and disgust. The unimodal approach showed 31.07% average recognition rate for the real-world dynamic AFEW data, 52.02% for RAF aligned data and 62.58% for CK+, a limited environment dataset. Moreover, 2% to 5% of data from natural settings went undetected, prompting the inclusion of an additional modality. By incorporating audio modality, emotion recognition in real-world contexts is enhanced, providing both relevant and supplementary information. It showed improvement in overall recognition of natural emotions. It also highlights that data must be analyzed in coarse to fine manner to appropriately recognized key areas that are responsible to give prominent features. Future performance can be boosted by exploring deep learning models to capture dynamic and detailed features from videos and by investigating various feature fusion techniques.




#### REFERENCES

- [1] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep Learning for Human Affect Recognition: Insights and New Developments," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524–543, Apr. 2021, doi: 10.1109/TAFFC.2018.2890471.
- [2] M. S. Hossain and G. Muhammad, "An Audio-Visual Emotion Recognition System Using Deep Learning Fusion for a Cognitive Wireless Framework," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 62–68, Jun. 2019, doi: 10.1109/MWC.2019.1800419.
- [3] T. Hassan *et al.*, "Automatic Detection of Pain from Facial Expressions: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1815–1831, Jun. 2021, doi: 10.1109/TPAMI.2019.2958341.
- [4] R. Srinivasan and A. M. Martinez, "Cross-Cultural and Cultural-Specific Production and Perception of Facial Expressions of Emotion in the Wild," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 707–721, Jul. 2021, doi: 10.1109/TAFFC.2018.2887267.
- [5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, IEEE, Jun. 2010, pp. 94–101. doi: 10.1109/CVPRW.2010.5543262.
- [6] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019, doi: 10.1109/TIP.2018.2868382.
- [7] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, DOI: 10.1145/2818346.2829994
- [8] D. Kollias and S. Zafeiriou, "Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition," 2018, [Online]. Available: <http://arxiv.org/abs/1811.07770>
- [9] J. Kossaifi *et al.*, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021, doi: 10.1109/TPAMI.2019.2944808.
- [10] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-Visual Emotion Recognition in Video Clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, Jan. 2019, doi: 10.1109/TAFFC.2017.2713783.
- [11] A. Birhala, C. N. Ristea, A. Radoi, and L. C. Dutu, "Temporal aggregation of audio-visual modalities for emotion recognition," in *2020 43rd International Conference on Telecommunications and Signal Processing, TSP 2020*, IEEE, Jul. 2020, pp. 305–308. doi: 10.1109/TSP49548.2020.9163474.
- [12] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, IEEE, Sep. 2019, pp. 552–558. doi: 10.1109/ACII.2019.8925444.
- [13] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric Learning-Based Multimodal Audio-Visual Emotion Recognition," *IEEE Multimedia*, vol. 27, no. 1, pp. 37–48, 2020, doi: 10.1109/MMUL.2019.2960219.
- [14] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Machine Vision and Applications*, vol. 30, no. 5, pp. 975–985, Jul. 2019, doi: 10.1007/s00138-018-0960-9.
- [15] M. K. Lee, D. Y. Choi, D. H. Kim, and B. C. Song, "Visual scene-aware hybrid neural network architecture for video-based facial expression recognition," *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, 2019, doi: 10.1109/FG.2019.8756551.
- [16] A. A. Salah, H. Kaya, and F. Gurnar, "Video-based emotion recognition in the wild," in *Multimodal Behavior Analysis in the Wild: Advances and Challenges.*, Elsevier, 2018, pp. 369–386. doi: 10.1016/B978-0-12-814601-9.00031-6.
- [17] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022, doi: 10.1109/TAFFC.2020.2981446.




- [18] W. C. De Melo, E. Granger, and A. Hadid, "A Deep Multiscale Spatiotemporal Network for Assessing Depression From Facial Dynamics," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1581–1592, Jul. 2022, doi: 10.1109/TAFFC.2020.3021755.
- [19] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, Apr. 2015, doi: 10.1016/j.neunet.2014.09.005.
- [20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.
- [21] K. Raut and S. Kulkarni, "Preprocessing Challenges for Real World Affect Recognition," *Computer Science & Engineering: An International Journal*, vol. 12, no. 6, pp. 43–52, Dec. 2022, doi: 10.5121/cseij.2022.12606.
- [22] K. Zvarevashe and O. Olugbara, "Ensemble learning of hybrid acoustic features for speech emotion recognition," *Algorithms*, vol. 13, no. 3, p. 70, Mar. 2020, doi: 10.3390/a13030070.
- [23] F. Haider and S. Luz, "Affect Recognition through Scalogram and Multi-resolution Cochleagram Features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, ISCA: ISCA*, Aug. 2021, pp. 581–585. doi: 10.21437/Interspeech.2021-1761.
- [24] A. Rushchyan and A. Khemchyan, "Human Emotion Recognition System," *International Journal of Engineering Trends and Technology*, vol. 72, no. 5, pp. 105–112, May 2024, doi: 10.14445/22315381/IJETT-V72I5P111.
- [25] N. George, L. Shine, A. N. B. Abraham, and S. Ramachandran, "A two-stage CNN model for the classification and severity analysis of retinal and choroidal diseases in OCT images," *International Journal of Intelligent Networks*, vol. 5, pp. 10–18, 2024, doi: 10.1016/j.ijin.2024.01.002.

## BIOGRAPHIES OF AUTHORS






**Karishma Raut**    has received a Master of Engineering from University of Mumbai and currently pursuing her Ph.D. in Electronics and Telecommunication Engineering from Bhartiya Vidya Bhavans' Sardar Patel Institute of Technology, Mumbai University. She has over 22 years of experience in the teaching profession and 4 years as Associate Professor at VIVA Institute of Technology, India. Her research interests include deep learning, affective computing, HCI, and related applications. She can be contacted at email: karishma.raut@spit.ac.in.



**Dr. Sujata Kulkarni**    holds Ph.D. in Electronics Engineering having work experience of more than 5 years in the industry. She has been in the teaching profession for over 16 years, of which she has been associated as a Professor with Bhartiya Vidya Bhavans' Sardar Patel Institute of Technology, Mumbai University for more than 5 years. Her areas of interest are image processing, machine learning, deep learning, and its applications. She can be contacted at email: sujata\_kulkarni@spit.ac.in.



**Dr. Ashwini Sawant**    holds a Ph.D. in Electronics and Telecommunication Engineering from Mumbai University. She has over 18 years of experience in the teaching profession, including more than 12 years as a Professor at Vivekanand Education Society's Institute of Technology, affiliated with Mumbai University. Her research interests encompass biomedical image processing, machine learning, pattern recognition, and its applications. She can be contacted at email: ashwini.sawant@ves.ac.in.