International Journal of

INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhanced Emotion Recognition for Women and Children Safety Prediction using Deep Network

Nanda R. Wagh[1], Dr. Sanjay R. Sutar[2]

Submitted: 27/05/2023 **Revised**: 07/07/2023 **Accepted**: 26/07/2023

Abstract: The most difficult research problem is ensuring the safety of women and children. Multimodal emotion recognition is a difficult task. One of the most important and widely used research domains in HCI is multimodal data, which includes audio, video, text, facial expression, body motions, bio-signals, and physiological data. This data is used to forecast the safety of women and children. Rigid research has been proposed in this context. To create the best multimodal model for emotion recognition combining picture, text, audio, and video modalities, a novel deep learning model is developed, and a thorough analysis of data, feature, and model-level fusion is undertaken. Separate innovative feature extractor networks are suggested specifically for picture, text, audio, and video data. Then, at the model level, an ideal multimodal emotion identification model is developed by combining information from images, text, voice, and video. Three benchmark multimodal datasets, IEMOCAP, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Surrey Audio-Visual Expressed Emotion (SAVEE), are used to evaluate the performances of the suggested models. On the IEMOCAP, SAVEE, and RAVDESS datasets, the suggested models obtain high predicted accuracies of 96%, 97%, and 97%, respectively. By contrasting the outcomes with those of the current emotion recognition models, the models' efficacy and optimality are also confirmed. Women and Children Safety Prediction employs multimodal Enhanced emotion recognition.

Keywords*: Facial Expression Recognition, deep learning, multimodal, women safety, audio-visual media, fusion*

1. Introduction

Human Their emotional moods have a huge impact on communication, which is quite important. It can be difficult to determine the emotional states in a variety of areas with a wide range of applications, including lie detection, audio-visual police work, affectional computing, online teaching and learning, online meetings, human-computer interface (HCI), and many, many more. The investigation of emotional variability is an essential component of psychological, mental adaptability, and well-being research. Additionally, machines need to be able to identify human emotions in order to make better judgements. The advancement of techniques to let intelligent computers accurately assess users' emotions is now essential for the progress of civilization since they must become an integral part of our daily lives.

Research commonly uses two models: dimensional models and discrete feeling models, despite the fact that there are numerous emotional models in the literature [4,5]. The former describes emotions as an infinite spectrum whereas the latter portrays them as unique values. Each model is regularly used by psychologists to evaluate emotions. Using voice data, physiological data, facial expressions, body language, and a variety of other elements, it is common practise to determine a person's moods. Each of those modalities has its own unique characteristics, thus combining them yields a chic depiction of options that can swiftly play-act feeling recognition. Wishing on a single modality for feeling recognition is futile, as has been shown in past experiments [6,][7].

Particularly, current research shows that using many modalities (audio, video, text, etc.) for emotion recognition yields considerably better results than using only one [8]. The former refers to emotions as a never-ending spectrum, whereas the latter shows them as discrete values. Each model is widely used by psychologists to investigate emotions. Multimodal period analysis of emotional states has received a great deal of attention recently from people all around the world under the aforementioned paradigm. Dynamic

[1] *Research Scholar, Dr. Babasaheb Ambedkar Technological University, Lonere (India), nandawagh@ dbatu.ac.in*

[2] *Professor and Head of Department, Dr. Babasaheb Ambedkar Technological University, Lonere (India), srsutar@dbatu.ac.in*

multimodal analysis of human emotions outperforms static analysis of human emotions because it takes into consideration elements like fluctuations in eye movements and facial expressions over time as well as voice characteristics [9]. Finding a cost-effective multimodal model in this field of research that is also not unduly complicated or computationally taxing is still a challenge. Therefore, it is anticipated that this work will be able to detect emotional states using a multimodal fusion model that considers both audio and video data. It is reasonably lightweight and reasonably priced. The two methods that have drawn the most interest in the last 10 years are multimodal emotion recognition (MER) and facial expression recognition (FER). FER will be produced using a variety of techniques, including face feature analysis, bi-signals, and the tried-and-true multi-modal approach [6],[3]. The various informational forms show the acknowledged outperformed lead.The modalities encompass all types of information, such as text cues, face images, acoustic expressions, linguistic data, semantic patterns, physical movements, eye gaze patterns, gestures, and electroencephalography (EEG) signals [2], [36], and [36] signals, as well as semantics like these signals. Numerous style strategies have been presented in recent research to automatically recognise affectional outcomes such as valence, arousal, dominance, and other emotion types [36, 37]. Emotions are the primary social communication components of importance. At some time in the future, depending on how it displays itself, our eyes will be able to tell whether it is okay or not.

Facial Features Recognition (FER) is a simpler problem. However, those who are deaf, visually disabled, or less sensitive are unable to convey their own sentiments. The most difficult FER evaluations are those performed by machines. By using the proper collection of algorithms and employing a multimodal strategy, the right set of emotions can be predicted. The necessity for girls' safety has increased, as it does in modern settings, and technology is crucial in modifying it. People often think of the face as their mental hub. As it makes different facial expressions, the face will produce a range of small Signals [56]. The term "emotional sensation" refers to the mental state that the human mind is capable of seeing and categorizing in order to appraise a range of appearances Interaction becomes more personalized and distinctive when computer algorithms attempt to grasp the user's purpose. The call from the computer becomes more significant as it starts to ask questions based on the user's mood. An autonomous vehicle might decide to travel further if the user's emotions are prone to anger, for instance. It is an excellent and extraordinarily sensitive indicator for figuring out how people behave, intend, think, and feel. With human safety and the aid of a feeling index, we frequently carry out autonomous tasks like marketing, observation, car safety, and appraisal. It will be simpler for people to communicate with technology in all imaginable ways thanks to the human-computer interface (HCI) industry. Interaction becomes more personalized and distinctive when computer algorithms attempt to grasp the user's purpose.The call from the machine becomes more important as it begins to pose queries based on the user's emotional state. An autonomous vehicle might decide to travel further if the user's emotions are prone to anger, for instance. It should start playing calming music when a user is feeling down. Young girls and kids can receive rapid support with pressing problems during the victimization phase thanks to the multimodal feeling identification technology, which assesses sentiments efficiently. Having outstanding, accurate, and acceptable judgement is crucial since good judgement is the capacity to discern between various emotions. A person's spirit can be effectively communicated through their face, voice, body language, gestures, movies, and even particular circumstances. Research of researchers shows that these methods attempt to accurately predict emotions to varied degrees.

To obtain higher precision, a variety of techniques and tactics have been discovered and are currently being applied worldwide. Paul Ekman, an associate degree creator, claimed that happiness, anger, surprise, sorrow, concern, and dislike are the six basic emotions that comprise a human emotional state. Ekman also noted the propensity to obfuscate human emotions by utilizing units of action (AU) [9]. play [1]. Numerous additional modalities eventually emerged as a result of the fact that facial features predominated among the modalities for identifying moods. The major objective of this paper is to do multimodal emotion recognition utilizing text, facial images, and sound. The video is immediately used as the source for the multimodalities.Girls and children are the targets of really severe violence. Violence is contagious in the society that is governed by men. It's too late to stop the breakout scenario now. How can crime against girls and children be addressed and reduced? By implementing positive society improvements, crime can be decreased and perhaps even managed to some extent. Molestation, robbery, and rape incidents happen frequently. Physically abusing girls and children as part of domestic violence is common. We treat every square foot like an animal. Most of the time, while counting, the simplest square measurement is used. The "WOMEN EMPOWERMENT" world is still a "fantasy" on a grand scale. Depressive disorders are more prevalent in girls and young people.

By using deep learning and machine learning to recognise multimodal emotions, the problem of women's and children's safety will be resolved. The technology paradigm will lessen the victimization of women and children. Concerns about the safety and wellbeing of women and children, particularly in public settings, have grown over the past several years. The necessity for rigorous protocols to secure their protection and give prompt aid has been underlined by incidents of harassment, assault, and abuse. In this situation, multimodal emotion identification technology has emerged as a possible solution. This technology uses a variety of sensory inputs, including body language, vocal intonation, and facial expressions, to precisely analyze and decode human emotions.

Multimodal emotion identification has the potential to considerably improve current safety precautions for women and children while also enabling people to react skillfully to looming risks. The purpose of this research paper is to study the idea of multimodal emotion identification and its applications in the context of women's and children's safety, emphasizing the importance of this idea in establishing safer environments for those who are more vulnerable. The goal of this project is to provide a way for employing convolution neural networks to recognize emotions related to the protection of girls, women, and children. When employing a digital camera to capture people's faces over the given time period, the model is utilized to identify their moods. This work's overall contributions are represented as follows:

1. The research contribution is predictive analysis of emotion based on multimodalities and utilized for women and children safety.
2. The proposed robust convolution neural network model building and decision level fusion for high accuracy.
3. Deep learning-based feature extractor networks for images, text, video, and audio.
4. A model-level fusion of the video associate degreed audio options is performed to make an optimum multimodal feeling recognition mode

The structure of part of this article is as follows: The second part presents the literature review, the challenges of multimodal emotion recognition and multimodal data sets. The third part focuses on transfer learning, hyper tuning, and novel convolution neural network model design with optimized parameters. The fourth evaluates the relevant experimental results and analyzes. The fifth part state summary, the conclusion and future scope.

2. Literature Review

II.1 Motivation

Literature Review

A person's emotional condition is indicated by the word "emotion." Emotional manifestations can be found in studies, passions, thoughts, and gestures. Stoic emotion, like Cicero, is divided into four categories: fear, pain, pleasure, and anger [7]. In a subsequent statement, Charles Darwin claimed that the late 19th century saw widespread acceptance of the evolutionary explanation of sentiments. The eight fundamental forms of sensations that researcher Plutchik [7,8] recently classified into are visually represented by the wheel of feelings in Figure 1. Ekman [9], [1][, 24] also made a disastrous connection between facial expression and emotion. ERC (Emotion Recognition in Communication) expresses the assessment of interactions in social networks.

It is beneficial for assessing in-the-moment communications that might be pertinent to court proceedings, interviews, security monitoring systems, e-health services, and other circumstances. [1],[9], [32]. Utilising 2D convolution neural networks (CNN), point birth and bracket enforcement [19]. The current emulsion method, which comprises of an early emulsion (functional position) and a late emulsion (decision position), is a cold-blooded emulsion method because it mixes the functions and opinions at various phases. [9],[1],[2],[15],[17],[21]. Ex.B. DialogueRNN, the most recent ERC projectNumerous important exploration problems are addressed and broken down by [1],[2],[11]. a two-model list of feelings. Feelings are defined using two models: category and dimensional. Emotions are categorised into a predetermined number of distinct orders in the category model.

In contrast, the dimensional model depicts emotion as a point in a continuous multidimensional space.[1],[2],[28],[22]. niceness Attention to detail Positive frame of mind Joy, alertness, rage, admiration, expectation, serenity, dread, and boredom; sadness, surprise, disgust, grief, awe, horror, and loathing of terror; and interest, affectation, acceptance. In view of the rapid development in the field of artificial intelligence, people are considering the emotional intelligence sector of robots (González Yubero, LazaroVisa, & Palomera, 2021).[34][35][2]. We are especially worried about the possibility that intimate human contact could be replicated by computer communication (Hanafi & Daud, 2021).

The most effective method of establishing computer-human communication is facial expression recognition, therefore this can only be accomplished if computers are able to record people's emotions.

The term "facial emotion recognition system" refers to a method of managing the psychosomatic feelings of the mortal demand for predictability by separating the precise facial drive of a stationary image from an evaluation of an active videotape sequence [12] [19],[2],[38]. But in nocturnal communication, face expression is an important and distinguishing manifestation. Changes in facial emotion can help us tell who is engaging in affective divagation [Furlong et al., 2021; Graumann et al., 2021; Staff etal., 2021]. Computer vision research is heavily focused on the detection of facial expressions. This can only be done if computers are able to accurately capture people's emotions since a computer vision recognition system will accurately distinguish face emotions during computer-human communication.

The domains of situation security backed driving, psychosomatic recognition, and many more will see a vertical increase as a result [20],[21]. In order to better categorise different emotions, we are modifying the Facial Emotion Recognition (FER) system (Hajarolasvadi and Demirel, December 2020; Rajananda, Zhu, and Peters, 2020). The seven orders of alienated mortal face expressions of joy, sadness, fear, wrath, surprise, anxiety, and neutrality. In order to identify facial expressions, the collected pictures must first be analysed with feature birth and bracket recognition. Since it has been made available, many scientists are likely to use CNN (Convolutional Neural Network), a well-known deep learning technique, to extract information from photos. A supervised literacy system that is comparable to an SVM system should be used, according to recommendations made by Ali, Hari Haran, Yaacob, and Adom (2015)[29][30]. Another approach offered by Evans (Evans, 2017) was the Haar surge transfigure (HWT) system. Phillips (Phillips, 2018) offered a unique system that combined two categories. The two styles are stationary sea entropy and Jaya's algorithm [34], [35], [2]. When we watch other people, we all become more tense.

According to a thorough examination of the materials [30], the issue is emotional face features that are eradicated with the styles. It is simple to forget if there is clear emotional recognition proof. These recognition models' broad description and power are customary in a similar way. Consequently, the facial emotion recognition model's efficacy is minimal. The model becomes less delicate [13]. There is a thorough investigation of both traditional and highly literate forms for multimodal emotion recognition. Due to the wide range of applications, it can be used in, the multimodal emotion bracket has attracted the interest of experimenters from all over the world. A significant amount of research is continually being conducted in this area. Researchers and psychologists have published a variety of methods for classifying emotions in the literature [12][35].

By Shaver et al., one of the most significant techniques for representing emotions was created. The authors broke emotions into prototypes to see if vibrant pieces of emotional data might be organised into a coherent whole. According to the authors' theory in [36], every emotion is distinct, tone-contained, and linked to every other emotion via a hierarchical form. The Ortony, Clore, and Collins (OCC) emotion model was created by Ortony et al. [12] and is based on the distribution of related emotions according to the thrill levels of those feelings. Less in-depth information about the excrescencies of the OCC model[10] was provided by Steunebrink et al. [37]. The OCC model was condensed to only include the two unique emotions of "nausea" and "interest." A further intriguing theory is offered by Ekman[15], in which emotions are seen to be quantifiable and biologically distinct. Six emotions were grouped together by Ekman: anger, surprise, grief, joy, nausea, and fear. The Ekmann model was transformed into a "Wheel of Feelings" model by Plutchik. As well as being introduced as emotion models, multi-dimensional models. A noteworthy example of a three-dimensional emotion model is the Latina cell model[22][16]. Through the use of the three confines, it can fit twenty various emotions into the three-dimensional mode. Eventually, Cambria et al. [14] presented the "Hourglass model," a novel emotion bracket model inspired by psychology and biology that is based on the models overall description and strength of these recognition models, which are also common. Therefore, the effectiveness of the model for recognising facial emotions is low. The model becomes less accurate [13]. This section presents a thorough analysis of conventional and deep learning-based approaches to multimodal emotion identification. Multimodal emotion categorization has attracted the interest of academics from all over the world due to its broad variety of applications, and a sizable quantity of research is carried out in this field each year.

A summary of emotion categorization models and algorithms is provided in Table 2.

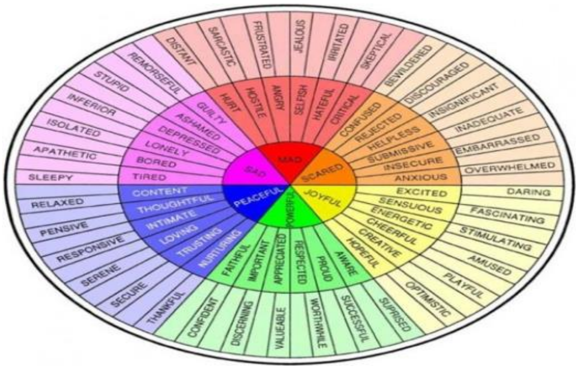


Fig. 1. Feeling wheel of emotion [36]

Table 2 A summary of emotion categorization models and algorithms.

Author	Models	#Emotions	Emotion type	Emotions
Shaver et al. [13]	–	6	Negative	Fear, anger, sadness.
Cambria et al. [14]	Hourglass model	24	Aptitude	Loathing, disgust, boredom, acceptance, trust, admiration.
			Sensitivity	Terror, fear, apprehension, annoyance, anger, rage.
			Attention	Amazement, surprise, distraction, interest, anticipation, vigilance
			Pleasantness	Grief, sadness, pensiveness, serenity, joy, ecstasy
P. Ekman [15]	P.Ekman Model	6	Negative	Disgust, fear, anger, sadness.
			Positive	Surprise, joy.
A. T. Latinjak [16]	Latinjak cube	20	Neutral	Fatigued, surprise, calm, alerted.
			Positive	Over-confident, relaxed, relieved, optimistic, satisfied, excited, enjoying, elated.
			Negative	Dejected, bored, sad, pessimistic, deceived, anxious, distressed, angry.
R. Plutchik [17]	Wheel of	8	Negative	Disgust, fear, anger, sadness.
	emotions		Positive	Anticipation, surprise, trust, joy.
Ortony et al. [12]	OCC	22	Negative	Pity, resentment, disappointment, fears-firmed, hate, reproach, shame, distress,
				fear.
			Positive	Gloating, happy, relief, satisfaction, gratitude, gratification, love, admiration, pride,
				joy, hope

Numerous studies use conventional techniques for multimodal emotion recognition. Busso et al. [10],[33] illustrated the advantages and disadvantages of an emotion identification system based on visual expression or audio input, as well as the use of two modalities in combination. The support vector machine was utilised to classify the emotional states using the audio-visual input, showing that the system outperformed both uni-modal systems. Overall classification accuracy for the feature-level fusion strategy was 89.1%, compared to 89.0% for the decision-level fusion approach. An technique to emotion detection employing audio-visual modalities was introduced by Wang et al. [58] [50] [52][32]. They used several speech features and a Gabor filter library to extract facial features. The stepwise technique was used to choose the relevant features for the classification of distinct emotions using Fisher's Linear Discriminant Analysis (FLDA) methodology. 82% total accuracy was attained by the suggested system. With the help of both face and speech features, Yan et al. [28] have presented a novel bimodal emotion recognition technique. Using a combination of human speech and facial expressions, Xu et al.'s [27] emotion recognition experiment. Eight emotions are contained in a dataset used by them called CHEAVD [38]. Face and speech features were recognised using SVM. Following that, Bayesian procedures were applied at the decision level to aggregate the classification results. For the CHEAVD dataset, their model's overall performance resulted in a 38% accuracy rate. The SVM model is used by Rao et al. [[50]23] to perform decision level fusion for the detection of various emotional states. Yoshitomi et al. [34] presented a multimodal strategy that takes into account not only auditory and visual data but also the temperature distributions recorded by an infrared camera. To combine these modalities at the decision-level, they used weights that had been discovered through experimentation. In many data science fields, deep learning techniques recently outperformed conventional ones.

This project involves the investigation, analysis, and development of a Machine Learning for Women and Children Safety System (WCSS) application based on various IoT and Machine Learning applications for women's and children's safety.[68] [69]. a useful device with a switch for weight. An assailant may squeeze the device by squeezing or packing it when they were about to assault the woman or child or when they noticed any signs of fragility in a more fascinating person.[66]. The success of these women's safety gadgets must be emulated for child safety devices given the rise in crimes against youngsters. However, child safety devices suffer from a substantial and unique challenge: the ability of the victim to interact and actuate the device in a conscious and timely fashion. Women safety devices are products and solutions targeted towards adult individuals who are conscious of their safety and can actuate an emergency device and actively deter the assaulter. However, children, especially those who are either not in their teenage, or are in the early stages of teenage, may not always be as cautious, conscious, and alert to actively keep track of their surroundings and engage defensive devices in a timely and efficient fashion. Regardless, substantial progress has been made in this direction and may researchers have proposed different approaches and devices to ensure the safety of children [66]. The Women and children Safety based App, Heat map-based location safety prediction can be performed. The surrounding effect of the shock absorption on the Convolutional Network is regarded as an important problem to the system. Zero insulation might not be the most optimal way of handling extremes, specifically on tiny images like heat maps, and it should be ignored when it comes to making because they barely translate to the actual world. Another solution to increasing the maximize objective is to use the asymmetric injection method and expand the number of the heatmap providing an even more optimized target because the last user is added[67].

Pleasantness	Attention	Sensitivity	Aptitude
Ecstasy	Vigilance	Rage	Admiration
Joy	Anticipation	Anger	Trust
Serenity	Interest	Annoyance	Acceptance
Pensiveness	Distraction	Apprehension	Boredom
Sadness	Surprise	Fear	Disgust
Grief	Amazement	Terror	Loathing

Table 1 Sentic levels of the Hourglass model [14][20]

As a result, it is no surprise that they have been adopted here as well. Nguyen et al. [25] introduced a novel method that integrated 3D CNN for 2.1 Multimodal Facial Emotion Recognition Challenges

There are multiple challenges for Multimodal emotion recognitions and listed as follows.

1. To establish Robust, Efficient, lightweight, and optimal deep learning multimodal model.
2. Analyzing variability of existing emotion datasets.
3. To determine the Affective domain value of emotion intensity.
4. To analysis and represent limitation of uni-modality.
5. To perform fusion at data level, feature level and decision level to achieve better accuracy.
6. Deaf, blind, and physically challenging, less sensitive person cannot express emotions. In real life Facial Expression Recognition (FER) is an easier task but for deaf, blind people and physically challenging, less sensitive people the fer is very difficult. Feature extraction is a crucial challenging task in multimodal approach [16].
7. Feature extraction is a dynamic process. Real-time data feature exact, emotion extractions are very problematic to achieve.

Fig 1 shows the extreme variety of emotions. The biggest hurdle about to recognize 78 classes of emotion with accuracy challenge for researcher. In Fig. 2, the categorical front, Plutchik's wheel[12] [8] defines eight types of discrete primary emotions, each of which has more subtle related subtypes. On the other hand, it is a great challenge to accurately predict the intensity of emotions [33],[36],[24].

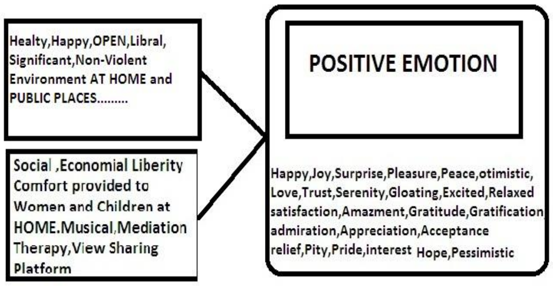

Healty,Happy,OPEN,Libral,
Significant,Non-Violent
Environment AT HOME and
PUBLIC PLACES.........
Social ,Economial Liberity
Comfort provided to
Women and Children at
HOME.Musical,Mediation
Therapy,View Sharing
Platform
POSITIVE EMOTION
Happy,Joy,Surprise,Pleasure,Peace,otimistic,
Love,Trust,Serenity,Gloating,Excited,Relaxed
satisfaction,Amazment,Gratitude,Gratification
admiration,Appreciation,Acceptance
relief,Pity,Pride,interest Hope,Pessimistic

Fig. 2 Shiv_Swar_Nanda Emotion model for Women and Children Safety

The figure 2 shows the Shiv_Swar_Nanda Emotion model for Women and Children Safety. If positive emotion is derived in real-time, women and children both are in safe mode. If the negative set of emotion is derived from multimodal dataset women and children both are unsafe mode. The unsafe mode means women and children has problem. The problem may be either victimization or any other personal probelm. Or else they may be suffering from some own some health problem which creates negative sets of emotion.

Research Challenges

There are many more challenges in data set, feature extraction. Innovative and unique way of feature extraction is always miserable. Hardware GPU costing and various research resources costing is substantial. Data set collection is also a viscous process. Challenge to get the exact set of notions classes for all modalities like text, speech, and images.

Training for the modalities is combination to improve the model accuracy for prediction is another challenging aspect to all researchers [26][27]. There are drastic changes in face expression in fraction of second. There is heavy confusion between sad and neutral. Shape of mouth, gesture, tone, and bio-signal change at dynamic rate. The processing of dynamic video sequence is also very complex. Machine learning computation challenges associated with it and many hurdle in gaining resources.

3. Methods

1.1. Datasets

Multimodal Datasets

IEMOCAP, SAVEE and RAVDESS are multimodal datasets used for experimentation.

Three benchmark multimodal databases, the IEMOCAP [3], SAVEE [2] and the RAVDESS [1] are used in this study.

IEMOCAP has 12 hours of audiovisual material on 10 actors, including discussion between an actor and an actress that was both scripted and spontaneously recorded [1],[3]. The audiovisual data is compiled into short sentences that last between three and fifteen seconds, which are subsequently identified by the assessors. Three to four different people analyse each statement. Ten options (neutral, happy, sad, angry, surprised, afraid, disgusted, frustrated, delighted, other) were provided on the evaluation form. We only examine four of them because that is what prior research has done: anger, excitement (happiness), neutrality, and sadness. Very Good Form According to earlier studies, we consider emotions when at least two experts concur with their choice,1],[10][15].

The Surrey Audio-Visual Expressed Emotion (SAVEE) database has been documented as a requirement for the creation of an automatic emotion recognition system. The database contains 480 British English utterances recorded from 4 male actors portraying 7 different emotions. The sentences were phonetically balanced for each mood and taken from the typical TIMIT corpus. High-end audio-visual equipment was used to record, process, and label the data in a visual media lab. Ten volunteers examined the recordings under auditory, visual, and audio-visual circumstances in order to assess the performance quality. For the auditory, visual, and audio-visual modalities, classification systems were created using standard features and classifiers, and speaker-independent identification rates of 61%, 65%, and 84% were obtained, respectively [2].

A multimodal dataset called RAVDESS [1][11] contains 7356 files totaling 24.8 GB. These statistics are from 24 professional actors (12 men and 12 women), who each spoke two lexically similar phrases with a North American accent. Disgust, surprise, fear, sadness, joy, anger, and calm emotions can all be heard in speech. Every display of emotion has two unique emotional intensity levels (strong and normal), as well as a neutral expression.

The data modalities included in this database are Audio-Video, Audio-Only, and Video-Only (without sound). Because the objective of this work is to perform multimodal emotion recognition, which requires the use of both audio and visual data for each actor, a fraction of the video speech files (i.e., files with both audio and visual modalities) are employed.

There are 1440 files in total, which are divided into eight emotion classes: fearful, disgust, angry, sad, happy, calm, neutral, and surprised.

Four researchers and students from the University of Surrey, ranging in age from 27 to 31, recorded the data for the SAVEE [2][10] dataset. This dataset contains 480 audio–visual files, with 120 utterances for each speaker. All the audio–visual files are in. avid format and there are seven emotion classes namely surprise, sadness, neutral, happiness, fear, disgust, and anger.

FER2013 database. The data collection used for the application was the FER2013 dataset from the Kaggle challenge on FER2013[38]. The database is used to incorporate the Facial Expression detection framework. The dataset consists of 35,887 pictures, split into 3589 experiments and 28,709 images of trains. The dataset includes another 3589 private test images for the final test.

Multimodal approach uses text, images, speech, and video sequences.

Enhancement for the anticipated CNN. A fine-tuning methodology replaces the pre-trained model's completely connected layers with a new set of completely connected layers when training a model on a given dataset. It also uses backpropagation to fine-tune all or some of the kernels in the pre-trained convolutional subcaste base. The active factors that control the affair size of the convolutional subcaste include padding, stride, batch size, sludge, sliding window, and literacy rate parameter. Padding is necessary to give the input border bottoms. The height and range parameters are distributed by Stride. Small stride lengths provide wide fields and massive, widely overlapping affairs.

The open fields lap lower, with longer strides and less limitation. While fine-tuning all of the layers in the convolutional base, it is possible to tune a considerable portion of the deeper layers. The proposed model in this work has four layers of complexity and two fully connected layers. In order to complete this task, just the completely linked layers that serve as a classifier and the detailed point block material at high position would need to be trained. Since we only had 7 sensations, which is a disparity, the SoftMax rating was reset to 7 grades from 1000 species. an anticipated CNN pipeline.

A network that recycles input using sub casting. The latter two levels are entirely connected to the last four layers of complexity and new pooling. Batch normalization, ReLU subcaste, and a completely linked subcaste of each of the four network topologies are used to handle any complexity. The new thick subcaste is used at the end of the four complication layers, which are connected to the two totally connected layers. One of the two possibilities serves as the foundation for the general channel in the proposed CNN model. Fig. 3. The basic mechanism for recognizing emotions is depicted in Fig. 4. On the face picture or modality, preprocessing and feature extraction are carried out. It is anticipated that human beings would be classified according to their performed positive and negative mood states.

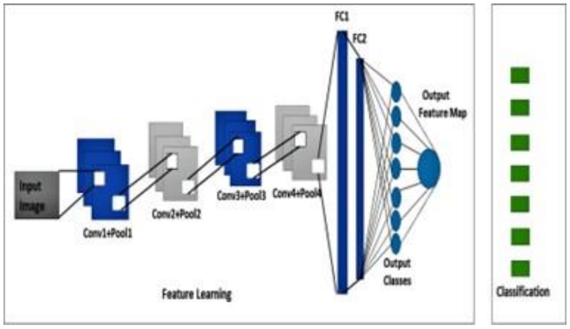


Fig. 3 Proposed Convolution Network Architecture

Data preprocessing and feature engineering

In this section we describe how we extracted our video and audio features and prepared our multimodal labeled dataset

Video preprocessing and feature engineering

It was determined that this was adequate to capture the spatiotemporal data related to a specific mood, therefore for each actor in each dataset, a total of 6 frames from each movie with a duration of 3 seconds are gathered. Be aware that the video frames are retrieved from the videos using the computer vision tool OpenCV (version 3.3) [54]. The frames are then carefully processed using OpenCV's built-in pre-trained deep learning-based face detection. In this experiment, a face detector constructed using Caffe is used. The.prototxt and.caffemodel files, which are downloaded and used during runtime, contain the model's architecture and the weights of each layer.

The deep learning-based face detector makes use of the Single Shot Detector (SSD) architecture. The extracted frames from each video are scaled back to 40% of their original size. The blobFromImage() function built within the DNN module is then used to create a blob from each image. The blobFromImage() function handles the necessary image preprocessing steps, such as blob dimension setting and RGB normalisation. After normalisation and mean pixel intensity removal, it creates a blob that represents the input image in four dimensions. Now that the blob has been subjected to the network's face recognition algorithm, the detection scores (probabilities) of the facial predictions have been identified. The detection scores are compared to the confidence level in order to exclude the subpar detections. After creating a bounding box along the facial region, the image's x-y coordinates are calculated using the detection scores. The image is then appropriately cropped. Afterwards, the image will be cropped, reduced in size to 64 64 pixels, and added to a list of features. In order to maintain them in the list of video features, aspects are removed from each video.

3.1 Multimodal Facial Emotion Recognition

Positive and negative emotion sets exist in human beings. For the safety of women and children, the proposed research uses a negative set of emotions. The automatic women and kid safety prediction model that is suggested. A significant innovation in emotion discrimination is introduced by the suggested approach. The negative emotions of anger, sadness, disgust, and fear can be utilised to forecast safe and unsafe situations based on threshold levels.

The application of machine learning for the safety of women and children centres on the multimodal method of emotion detection that includes audio, video, text, and facial expression recognition. For both men and women, we looked at experimental studies on multimodal multimodalities combining features that often perform better than uni-modal features. Utilizing multimodal over unimodal resulted in greater relative improvement.

The most effective multimodal classifier for both genders is one that combines text, audio, and visual information. Emotion detection is improved and is usually more accurate when the emotions produced from many models are combined. MT-CNN (Multitask-Convolution Neural Network) and Convolution Neural Network with Hyper Parameters are two examples of deep learning techniques that can be used for image emotion detection. FER, with bespoke training, is also used to improve the recognition of face expressions of emotion.

Deep learning for vocal emotion recognition and multimodal emotion recognition in the IEMOCAP data set Using a CNN, text sentiment analysis and bidirectional LSTM emotion recognition.

According to the plan, CNN pulls the local characteristics from the sentence's context once the LSTM layer has collected it. employing a rule-based technique to analyse pulse rates. Combining the two approaches may lead to a better comprehension of emotions and more precise emotion prediction.

We suggest a special remedy for the current FER issues. Transfer learning and hyper-parameter tuning are the foundations of the heightened face emotion recognition novel model. We make use of the ResNet-50 and Dense Net networks. Using the most recent deep learning methodology, auto feature extraction is conducted. the activation and maximum pooling convolutional neural network. It employs the ReLU and Softmax optimizers. The model advances beyond the convergence. On the Google Colab GPU, the experiment was run. The total rate of model recognition increased. The combination of the suggested method is superior to the most sophisticated way. 95.91 percent of the training was accurate. The authors propose a CNN technique for identifying seven facial expressions and real-time facial expression recognition using the merging of Convolutional Neural Network (CNN), Local Binary Pattern (LBP) features, and Oriented FAST and rotating BRIEF (ORB).

We provide a four-layer "Convolution Neural Network" model for this system that makes the greatest use of CNN optimized parameters.

In order to improve generalization and the precision of learning and prediction, this research exposes the aggregate photos from several databases.

Longer training sets rather than training and testing sets might produce results that represent higher consistency. It can also include improved testing methodologies, such as preparation, testing, and verifying processes.

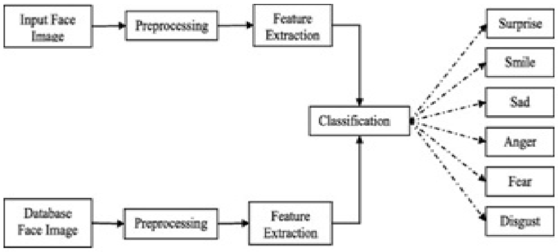


Fig. 4. Fundamental Emotion Recognition System

Large, medium, and small datasets are used to assess how well the "ConvNet" architecture performs. The results show that our technology is capable of performing superbly in both cases.

In a small number of epochs, this work achieves a training accuracy of over 97%, demonstrating that the model is well calibrated for the method. The generalization approaches' classification accuracy on the IEMOCAP, SAVEE, and RAVDESS datasets was 96%, 97%, and 97%, respectively.The research contribution is predictive analysis of emotion based on multimodalities and used for women and children safety. The proposed robust convolution neural network model and decision level fusion for high accuracy.

3.1.1 Data preprocessing and feature engineering

In this section we describe how we extracted our video and audio features and prepared our multimodal labeled dataset.

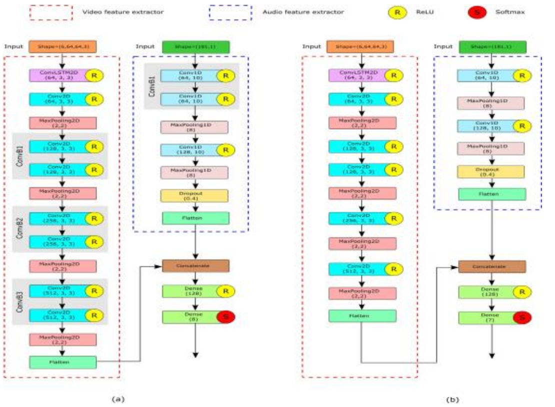


Fig. 5. Video and Audio Feature Extractor

3.1.1.1 Video preprocessing and feature engineering

Six frames per videotape (length: 3 s) are used for each actor in each dataset since it is thought that this number of frames will be adequate to capture the spatiotemporal information connected to a particular mood. Be advised that the films' frames were removed using the machine vision software OpenCV(interpretation3.3)(54). OpenCV's built-in, deeply rooted, pre-trained face sensor is also used for exact facial point birth from the frames. A Caffe- grounded facial sensor is employed in this experiment. The model's armature and the weights for each subcaste are stored in separate prototxt and caffemodel lines, which are downloaded and used during runtime. The highly literate facial sensor makes use of the Single Shot Sensor (SSD) frame.

After being rooted, the frame size of each recording is decreased by 40%. Each image is additionally supplied by the blobFromImage() function, which is a part of the DNN module. The blobFromImage() function takes care of the necessary image preparation, which includes things like setting the blob confines and doing RGB normalisation. It returns a four-dimensional blob that is a representation of the input image after normalisation and mean pixel intensity deduction. As the blob moves through the network, faces will be found using the discovery scores (chances) of the facial prognostications that have been put up to do so. By comparing the discovery scores to the confidence threshold, weak discoveries are filtered away. After building a bounding box along the facial region and calculating the box's x-y equals using the discovery scores, the image is then cropped. The cropped image needs to be resized to 64 by 64 pixels in order to be added to a list of features. In this manner, the face features from every videotape are removed and saved in the list of features for each videotape. Multiple Expressions of the Face are Recognised People who are mortal have both positive and negative emotion sets. In order to protect women and children, the proposed exploration makes use of a negative set of emotions. The foretelling of a system that automatically protects women and children. The suggested paradigm incorporates an essential innovation in the demarcation of sentiments. The negative emotions of anger, sadness, and fear can be utilised to predict safe and risky situations based on threshold levels. The proposed investigation for use of machine learning for the safety of women and children centres on the multimodal approach including audio, video, text, and facial expression discovery. We looked at experimental research on multimodal multimodalities coupled features, which typically perform better than single-modal features. It was established that using many modes is preferable to using just one. In order to create a multimodal classifier that performs well for both genders, visual, auditory, and text characteristics were combined. Combining the sensations inferred from many models leads to improved emotion discovery and is generally more accurate. Using a deep learning approach akin to the MT (Multitask-complicated Neural Network), image emotion discovery is achieved by combining FER with customised training to improve the recognition of facial emotions and numerous publicly accessible datasets.

Voice and emotion recognition in the IEMOCAP data set with deep literacy, multimodal emotion recognition Text sentiment analysis and bidirectional LSTM emotion recognition using a CNN 4 (Abecedarian Emotion Recognition System Convolutional Neural Network) subcaste. The concept is that the CNN extracts the original qualities from the LSTM subcaste, which records the data about the judgment's environment. Using both strategies together may result in a more accurate prognosis of feelings and a better comprehension of the sentimentsWe suggest a special solution to the FER difficulties. The novel model for improved face emotion identification is based on hyperactive parameter adjustment and transfer literacy. ResNet-50 and thick Net are the networks that we employ. bus point birth performed by rearmost deep literacy paradigm. the convolutional neural network with the best pooling and activation capabilities. The ReLU and Softmax optimizers are employed. The model advances over the convergence. The Google Colab GPU was used for the testing.The model's overall rate of recognition improved. The training accuracy is 96.91 percent. The authors propose a CNN system that combines a convolutional neural network (CNN), original double pattern (LBP) features, well-known FAST and rotational BRIEF (spherical) features, and can detect seven different facial expressions in real time.

For this system, we suggest a four-subcaste "Hybrid ConvNet" model with the chic application of CNN-optimized parameters. The collaborative photos from many databases are exposed in this investigation, which serves to improve conception and the delicacy of literacy and vaticination. In addition, it can produce results that show less thickness by using longer training sets instead of training and testing sets. It can also feature improved testing methods, similar to medication, testing, and validating processes. On big, medium, and small datasets, the effectiveness of the " ConvNet " armature is estimated. The results indicate that our system may achieve great performance in both scenarios. In a minimum number of ages, this study produces a training d accuracy of above 97, demonstrating the model's successful system acclimatisation. The accuracy obtained by the conception methods using the three datasets is, separately, 92.05, 93.05, and 95.13. The investigation is a multimodal, prophetic analysis of emotion used to ensure the protection of women and children. The decision position emulsion and strong ConvNet model that are suggested for high accuracy.

3.1.1.2 Audio preprocessing and feature engineering

Every video has its audio material extracted using the moviepy Python package. Since the retrieved audio was discovered to be a stereo file, it was split into a mono file using the pydub module [55]. then the mono audio file is used for feature extraction. Numerous audio features are retrieved and kept in the audio feature list, in particular MFCC, melspec- trogram, spectral difference, and tonnetz. Below is a list of the significance of each of those choices..

3.2 Transfer Learning, Multimodal Fusion

A key idea in deep learning is transfer learning. It employs the ideas of reuse and makes use of models that have been trained to tackle one problem as a jumping off point for another related issue. The most adaptable method of learning is transfer learning, which enables previously trained models to be utilised directly as feature extraction preprocessing and merged into whole new models [14] [18][15]. Many potent ImageNet image recognition task models, including VGG, Inception, and ResNet, are easily accessible using Kera's. The pre-trained traditional models for image categorization are ImageNet MobileNet, MobileNetV2, Xception, VGG16, VGG19, ResNet50, ResNet101, ResNet101V2, ResNet152V2, InceptionV3, InceptionResNetV2, and DenseNet. As seen above, ImageNet's annual Large Scale Visual Recognition Challenge (ILSVRC) featured a variety of high-performance picture categorization algorithms. Due to the image source utilised in the competition, this task is frequently referred to as simply "ImageNet," and it led to a number of advancements in convolutional neural network architecture and training. The foundation for transfer learning in machine vision applications can be efficiently employed with any model. There are a lot of admirable uses for it, including

Features that were learned and are useful As they were trained on more than 1,000,000 photographs over 1,000 categories, the models learned to recognise the general properties of the images. Modern Performance: The models showed modern performance and excelled in the particular image emotion detection job for which they were created.Simple to Access: Many libraries offer straightforward APIs for downloading and using models, and model weights are available as free downloads. Several alternative deep learning packages, including Keras [37][39], allow for the downloading and usage of model weights in the same model architecture.4. Experimentation and results

This section presents the experimental setup and result analysis for the approaches taken to train multimodal emotion detection. Following are the approaches:

1. **For training model using CNN, for IEMOCAP Dataset, the experiment results are as below:**

Epoch: 200
Batch: 228

```
Training Results
Epoch 200/200
228/228 [==============================] 18s
80ms/step
loss: 0.1391
accuracy: 0.9691
val_loss: 1.7726
val_accuracy: 0.6663
```

2. Deep neural framework - Conversational Memory Network for Emotion Recognition in Audio -Video

For CMN approach with Tensorflow 2.0 using GPU for training model.

Class labels: ('hap':0, 'sad':1, 'neu':2, 'ang':3, 'exc':4, 'fru':5)
Hyper parameter Tuning:

Parameter	Value	Description
mode	Text/Audio/Video	which modality
context	False	which kind of features to choose
nonlin_func	tf.nn.tanh	type of nonlinearity
learning_rate	0.001	Learning rate for SGD.
anneal_rate	60	Number of epochs between halving the learning rate.
anneal_stop_epoch	100	Epoch number to end annealed lr schedule.
max_grad_norm	40.0	Clip gradients to this norm.
Parameter	Value	Description
evaluation_interval	1	Evaluate and print results every x epochs
batch_size	512	Batch size for training.
hops	3	Number of hops in the Memory Network.
epochs	10	Number of epochs to train for.
embedding_size	100	Embedding size for embedding matrices.
input_dims	None	Number of timesteps of the RNN
timesteps	40	Number of timesteps of the RNN
class_size	None	No. of output classes
nonlin	True	Use nonlinearity
dropout_keep_prob	0.3	Percentage of input to keep in dropout

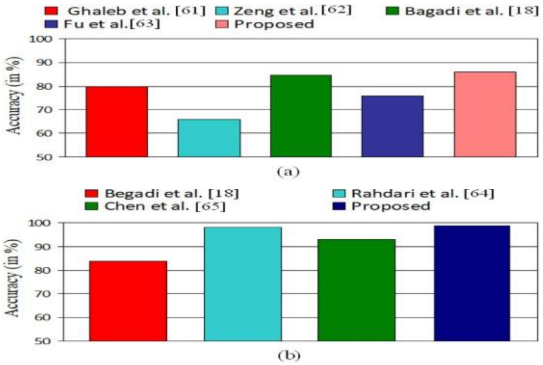
Training Results:
Epoch 10
Total Cost: 2.9286814853549004
Training Accuracy: 0.9795869191049914 Validation Accuracy: 0.5704008221993834 Validation Loss: 2.44280743598938.

To evaluate the effectiveness of the planned multimodal models, this section applies rigorous analysis. The Adam optimizer [59] with a learning rate of 0.0005 and a batch size of 64 is used to enforce all the models. One hundred epochs in total were considered for the model coaching. Eighty-third, 10%, and 100% information, respectively, are contained in the training, validation, and examination sets. The Google Colaboratory notebooks used for all the experiments have an Intel (R) Xeon (R) processor clocked at 2.00 GHz, 39 MB of cache, and an NVIDIA Tesla T4 GPU with 13 GB of useable memory (CUDA version: 11.2). The models are implemented in Python, which produces the best accuracy for the RAVDESS dataset of 0.86. Other effective combinations of audio and video feature extractors do, however, exist for this dataset, including V2 A3 (acc zero.77), V5 A3 (acc zero.76), V6 A5 (acc zero.76), V9 A3 (acc zero.76), V7 A5 (acc zero.76), V2 A5 (acc zero.76), V1 A1 (0.76), V2 A1 (0.75), V4 A1 (0.75), V9 A1 (acc zero. For the SAVEE dataset, experiments have also been conducted with different combinations of audio and video feature extractors.

The main goal of this study is to evaluate the performance of the combination of audio and video feature extractors used in RAVDESS, i.e., to determine whether or not a general model or dataset-specific model exists. It is found that, when compared to the other combinations, mix V2 A2 produces the best accuracy (acc 0.99). Nevertheless, similar to RAVDESS, there are various effective combinations of audio and video feature extractors for SAVEE, such as V1 A4 (acc zero.98), V2 A4 (acc0.98), V4 A2 (acc 0.98), V2 A6 (acc0.96), V4 A5 (acc 0.96), etc..

Detailled performance of the planned models (i.e., model with V8 A4 for RAVDESS and V2 A2 for SAVEE) along with additional arbitrarily selected best-performing combinations of audio and video feature extractors. Several performance metrics, including accuracy, specificity, recall, precision, AUC, and F1-score, are used to assess the success of the proposed models.

.



for (a) RAVDEES (b) SAVEE.

Fig. 6. Performance comparison of proposed models with the previous works

5. Conclusion

In this research article, a multimodal facial emotion recognition system is extensively investigated, and a multimodal convolutional network is proposed. The research emphasizes transfer learning, hyper parameter tuning, and the facial emotion recognition learning process. We have used the transfer learning concept of the popular current neural convolution algorithm combined with MobileNet50, which has had a worthy performance for effect on the multi-class classification. While performing validation on the data set, the experimental evaluation result has a good exactness and a good recognition effect in terms of average recognition precision. In future research, we will focus on exploring diverse facial emotion detection and will try to collect more real time emotional images, video, and transcripts than in this experiment in order to optimize and suggest a better algorithm to train the hyper parameters of the multilayer feedback neural network, such as weights and bias. We will also evaluate optimization algorithms based on the previous method to increase the performance of a multilayer feed forward neural network. We will continue to search for shapes based on a deep residual network to improve the accuracy of facial expression recognition. To develop robust deep learning model for emotion recognition of physically challenged, blind, deaf, and mentally retired human beings and old age person is future work of this research.

Compliance with Ethical Standards Disclosure of

6. Data Availability

I confirm that I or my relative have NO financial or other interest in the subject/matter of the work in which I will be involved, which may be considered as constituting a real, potential, or apparent conflict of interest. The data that support the findings of this study are available from the corresponding author upon reasonable request.
Research involving human participants and/or animals.
The proposed research review does not involve any human and animal participant.

Informed consent

As the proposed research review does not involve any human and animal participant no informed consent required from human or animal.

Acknowledgment

We are extremely thankful to Dr. Abhay E. Wagh, Director, Department of Technical Education, Government of Maharashtra; we are most deeply thanks to Dr. Santosh T. Yadav for his consistent motivation and lifetime support for research. We are very much thankful to Dr. Sanjay Nalbalwar, Head of the Department of Electronics and Telecommunication, Dr. Babasaheb Ambedkar Technological University Lonere. We thanks Dr. Brijesh Iyer, Assistant Professor, Department Electronics and Telecommunication, DBATU Lonere. We thanks a lot Dr. D.R. Nandanwar, Principal, Government Polytechnic Awasari for his moral and coherent support. Mr. Vishal Supekar for his unconditional and great support forever. We thanks Dr.Vijay Kohle and Mrs. Seema Kolhe for their worthy and inspirational support at step of life. Thanks Dr.S.M.Jadhav,Dr. Kiwalekar, Dr. L.D.Netak and Dr. V.J. Kadam for their significant guidance and consistent support. I thank to former head of Department Mr. U.V. Kokate,Government Polytechnic,Pune for his guidance. I thanks Mrs. J.R.Hange, Head of Department, Computer Engineering, Government Polytechnic Awasari for her support. I thanks all my colleagues form Government Polytechnic Pune and Government Polytechnic Awasari.

References

[1] Puri, T., Soni, M., Dhiman, G., Ibrahim Khalaf, O. and Raza Khan, I., 2022. Detection of emotion of speech for RAVDESS audio using hybrid convolution neural network. Journal of Healthcare Engineering, 2022.

[2] Singh, P., Srivastava, R., Rana, K.P.S. and Kumar, V., 2021. A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowledge-Based Systems, 229, p.107316.

[3] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42, pp.335-359.

[4] P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al., Combining modal- ity specific deep neural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 543–550.

[5] N. Srivastava, R. Salakhutdinov, et al., Multimodal learning with deep Boltzmann machines, in: NIPS, Vol. 1, Citeseer, 2012, p. 2.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, Multimodal deep learn- ing, in: International Conference on Machine Learning (ICML), Bellevue, WA, 2011, pp. 689–696.

[7] Y. Wang, L. Guan, Recognizing human emotional state from audiovisual signals, IEEE Trans. Multimed. 10 (5) (2008) 936–946.

[8] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee,

[9] U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in: Proceedings of the 6th International Conference on Multimodal Interfaces, 2004, pp. 205–211.

[10] Y. Yoshitomi, S.-I. Kim, T. Kawano, T. Kilazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in: Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499), IEEE, 2000, pp. 178–183.

[11] Z. Wang, S.-B. Ho, E. Cambria, A review of emotion sensing: categoriza- tion models and algorithms, Multimedia Tools Appl. 79 (47–48) (2020) 35553–35582, http://dx.doi.org/10.1007/s11042-019-

08328-z.

[12] A. Ortony, T.J. Turner, What's basic about basic emotions? Psychol. Rev. 97 (3) (1990) 315.

[13] B.R. Steunebrink, M. Dastani, J.J.C. Meyer, The OCC model revisited, in: Proceedings of the 4th Workshop on Emotion and Computing, 2009.

[14] Y. Li, J. Tao, L. Chao, W. Bao, Y. Liu, CHEAVD: a Chinese natural emotional audio–visual database, J. Ambient Intell. Humaniz. Comput. 8 (6) (2017) 913–924.

[15] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE' 05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006, p. 8, http://dx.doi.org/10.1109/ICDEW.2006. 145.

[16] E. Patterson, S. Gurbuz, Z. Tufekci, J. Gowdy, CUAVE: A new audio-visual database for multimodal human-computer interface research, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2002, pp. II–2017–II–2020, http://dx.doi.org/10.1109/ICASSP.2002. 5745028.

[17] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.

[18] A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon, Emotion recognition in the wild challenge 2013, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 509–516.

[19] J. Huang, J. Tao, B. Liu, Z. Lian, M. Niu, Multimodal transformer fusion for continuous emotion recognition, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, http://dx.doi.org/10.1109/icassp40776.2020.9053762.

[20] J.-B. Delbrouck, N. Tits, M. Brousmiche, S. Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, 2020, arXiv preprint arXiv:2006.15955.

[21] Y. Susanto, A.G. Livingstone, B.C. Ng, E. Cambria, The hourglass model revisited, IEEE Intell. Syst. 35 (5) (2020) 96–102.

[22] R. Plutchik, The nature of emotions: Human emotions have deep evolu- tionary roots, a fact that may explain their complexity and provide tools for clinical practice, Am. Sci. 89 (4) (2001) 344–350.

[23] K.R. Bagadi, A comprehensive analysis of multimodal speech emotion recognition, in: Journal of Physics: Conference Series, Vol. 1917, IOP Publishing, 2021, 012009.

[24] M. Abdullah, M. Ahmad, D. Han, Facial expression recognition in videos: An CNN-LSTM based model for video classification, in: 2020 International Conference on Electronics, Information, and Communication (ICEIC), IEEE, 2020, pp. 1–3.

[25] D. Krishna, A. Patil, Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks, in: Interspeech, 2020, pp. 4243–4247.

[26] A. Jaratrotkamjorn, A. Choksuriwong, Bimodal emotion recognition using deep belief network, in: 2019 23rd International Computer Science and Engineering Conference (ICSEC), IEEE, 2019, pp. 103–109.

[27] G. Sahu, Multimodal speech emotion recognition and ambiguity resolution, 2019, arXiv preprint arXiv:1904.06022.

[28] K.P. Rao, M.C.S. Rao, N.H. Chowdary, An integrated approach to emotion recognition and gender classification, J. Vis. Commun. Image Represent. 60 (2019) 339–345.

[29] S. Yoon, S. Byun, K. Jung, Multimodal speech emotion recognition using audio and text, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 112–118.

[30] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, C. Fookes, Deep spatio- temporal feature fusion with compact bilinear pooling for multimodal emotion recognition, Comput. Vis. Image Underst. 174 (2018) 33–42.

[31] H. Miao, Y. Zhang, W. Li, H. Zhang, D. Wang, S. Feng, Chinese multimodal emotion recognition in deep and traditional machine leaming approaches, in: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), IEEE, 2018, pp. 1–6.

[32] F. Xu, Z. Wang, Emotion recognition research based on integration of facial expression and voice, in: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE, 2018, pp. 1–6.

[33] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, B. Wang, Sparse kernel reduced- rank regression for bimodal emotion recognition from facial expression and speech, IEEE Trans. Multimed. 18 (7) (2016) 1319–1329.

[34] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic,

[35] P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al., Combining modal- ity specific deep neural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 543–550.

[36] N. Srivastava, R. Salakhutdinov, et al., Multimodal learning with deep Boltzmann machines, in: NIPS, Vol. 1, Citeseer, 2012, p. 2.

[37] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Ng, Multimodal deep learn- ing, in: International Conference on Machine Learning (ICML), Bellevue, WA, 2011, pp. 689–696.

[38] Y. Wang, L. Guan, Recognizing human emotional state from audiovisual signals, IEEE Trans. Multimed. 10 (5) (2008) 936–946.

[37] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in: Proceedings of the 6th International Conference on Multimodal Interfaces, 2004, pp. 205–211.

[39] Y. Yoshitomi, S.-I. Kim, T. Kawano, T. Kilazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in: Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499), IEEE, 2000, pp. 178–183.

[40] Z. Wang, S.-B. Ho, E. Cambria, A review of emotion sensing: categoriza- tion models and algorithms, Multimedia Tools Appl. 79 (47–48) (2020) 35553–35582, http://dx.doi.org/10.1007/s11042-019-08328-z.

[41] A. Ortony, T.J. Turner, What's basic about basic emotions? Psychol. Rev. 97 (3) (1990) 315.

[42] B.R. Steunebrink, M. Dastani, J.J.C. Meyer, The OCC model revisited, in: Proceedings of the 4th Workshop on Emotion and Computing, 2009.

[43] Y. Li, J. Tao, L. Chao, W. Bao, Y. Liu, CHEAVD: a Chinese natural emotional audio–visual database, J. Ambient Intell. Humaniz. Comput. 8 (6) (2017) 913–924.

[44] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE' 05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006, p. 8, http://dx.doi.org/10.1109/ICDEW.2006. 145.

[45] E. Patterson, S. Gurbuz, Z. Tufekci, J. Gowdy, CUAVE: A new audio-visual database for multimodal human-computer interface research, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2002, pp. II–2017–II–2020, http://dx.doi.org/10.1109/ICASSP.2002. 5745028.

[46] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.

[47] A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon, Emotion recognition in the wild challenge 2013, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 509–516.

[48] J. Huang, J. Tao, B. Liu, Z. Lian, M. Niu, Multimodal transformer fusion for continuous emotion recognition, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, http://dx.doi.org/10.1109/icassp40776.2020.9053762.

[49] J.-B. Delbrouck, N. Tits, M. Brousmiche, S. Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, 2020, arXiv preprint arXiv:2006.15955.

[50] Y. Susanto, A.G. Livingstone, B.C. Ng, E. Cambria, The hourglass model revisited, IEEE Intell. Syst. 35 (5) (2020) 96–102.

[51] Y. Soon, S.N. Koh, C.K. Yeo, Noisy speech enhancement using discrete cosine transform, Speech Commun. 24 (3) (1998) 249–257.

[52] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, L.-H. Cai, Music type classification by spectral contrast feature, in: Proceedings. IEEE International Conference on Multimedia and Expo, Vol. 1, IEEE, 2002, pp. 113–116.

[53] Y. Li, J. Yang, J. Wen, Entropy-based redundancy analysis and information screening, Digit. Commun. Netw. (2021) http://dx.doi.org/10.1016/j.dcan. 2021.12.001.

[54] I. Kandel, M. Castelli, The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset, ICT Express 6 (4) (2020) 312–315.

[55] E. Ghaleb, M. Popa, 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), IEEE, 2019, pp. 552–558.

[56] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools Appl. 78 (3) (2019) 3705–3722.

[57] Z. Fu, F. Liu, H. Wang, J. Qi, X. Fu, A. Zhou, Z. Li, A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition, 2021, arXiv:2111.02172.

[58] F. Rahdari, E. Rashedi, M. Eftekhari, A multimodal emotion recognition system using facial landmark analysis, Iran. J. Sci. Technol. Trans. Electr. Eng. 43 (1) (2019) 171–189.

[59] L. Chen, K. Wang, M. Wu, W. Pedrycz, K. Hirota, K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition, IFAC-PapersOnLine 53 (2) (2020) 10250–10254., http://dx.doi.org/10.1109/mis.2021.3062200.

[60] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-time video emotion recognition based on reinforcement learning and domain knowledge, IEEE Trans. Circuits Syst. Video Technol. (2021) 1, http://dx.doi.org/10.1109/ tcsvt.2021.3072412.

[61] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[62] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception- resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[63] OpenCV, Open Source Computer Vision Library, 2015.

[64] Priya Metri, Jayshree Ghorpade, Ayesha Butalia," Facial Emotion Recognition Using Context Based Multimodal Approach", International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 1, NO.4

[65] Nanda R. Wagh, Dr. Sanjay R. Sutar and Dr. Abhay E. Wagh," A Survey on the Recent Advances in the Development of IoT-based Devices for Women Safety", Intelligent Computing in Information Technology for Engineering System Proceedings of the International Conference on Intelligent Computing in Information Technology for Engineering System (ICICITES-2021), 25-26 June, 2021,https://www.routledge.com/Intelligent-Computing-in-Information-Technology-for-Engineering-System/Karande-Deshmukh-Mahalle/p/book/9781032270807.

[66] Nanda R. Wagh, Dr. Sanjay R.Sutar," An enhanced security of women and children using machine learning and data mining techniques", Data Mining and Machine Learning Applications,423-446,28 Jan,2022 https://doi.org/10.1002/9781119792529.ch16

[67] Nanda R. Wagh, Dr. Sanjay R.Sutar ,"A Smart Security Solution for Women{'}s and Children Using Wearable Iot Systems",Preprint SSRN, ISSN: 15565068

[68] AGYEI , I. T. . (2021). Simulating HRM Technology Operations in Contemporary Retailing . International Journal of New Practices in Management and Engineering, 10(02), 10–14. https://doi.org/10.17762/ijnpme.v10i02.132

[69] Anupong, W., Azhagumurugan, R., Sahay, K. B., Dhabliya, D., Kumar, R., & Vijendra Babu, D. (2022). Towards a high precision in AMI-based smart meters and new technologies in the smart grid. Sustainable Computing: Informatics and Systems, 35 doi:10.1016/j.suscom.2022.100690

Name of Author

Nanda R. Wagh completed the master's in computer science and engineering from Shree Ramanand Tirth Marathwada University Nanded, MS, India in 2009. Since then, she has worked as Assistant professor in the Department of Computer Engineering/IT at MIT Alandi,Savitribai Phule Pune University where his research interests include facial recognition, physical human-computer interaction, multisensory data fusion, multimodal emotion recognition, and women's and children safety. She has written a book on Artificial Intelligence for Anna University. She is currently working as Lecturer in Computer Engineering Department, Government Polytechnic Awasari under the Department of Technical Education, Mumbai. She is working as a research scholar at Information Technology, DBATU, Lonere.

Dr. Sanjay R. Sutar received Ph.D from Shree Ramanand Tirth Marathwada University Nanded, MS, India. He completed the master's in computer science and Engineering from DBATU, Lonere and B.Tech from Walchand College of Engineering, Sangali. Since then, he has working as Professor and Head in the Department of Information Technology at Dr. Babasaheb Ambedkar Technological University, Lonere, MS, India where his research interests include scheduling and Evolutionary Algorithm.