IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

ASSESSMENT OF THE DEEP CNN MODEL FOR SPEECH AND FACIAL EXPRESSION EMOTION RECOGNITION BASED ON THE MFCC

[1]K.Venkata Nagendra, [2]N. Harish , [3] Mallishetty. Praveen Kuamr , [4]P.Bhargavi

[1,2,3]Dept of Computer Science and Engineering, Sree Venkateswara College Of Engineering, Nellore (Dt), Andhra Pradesh, India.

[4]Dept of Electronics and Communication Engineering, Sree Venkateswara College Of Engineering, Nellore (Dt), Andhra Pradesh, India.

ABSTRACT

A huge development is done in current living in the disciplines of artificial intelligence, machine learning, human-machine interaction, etc. It is becoming more and more usual to interact with machines or give them instructions for specific tasks using voice commands. Siri, Alexa, Cortana, Google Assist, etc. are embedded into a lot of consumer electronics. But machines are constrained because they can't communicate with people the same way that people can. It cannot comprehend or respond to human emotions. Research in recognising emotions from speech is being led by the field of human-machine interaction. A more robust man-machine communication system is necessary given the significance of machines in our everyday lives. The goal of speech emotion recognition (SER), which is being developed by several researchers, is to improve communication between humans and machines. To do this, a machine must be able to identify emotional states and respond to them in a manner similar to how we humans do. The calibre of the retrieved features and the kind of classifiers employed determine how efficient the SER system is. The four main emotions—anger, sadness, neutrality, and happiness—from speech were the focus of this investigation. In this study, distinct emotions are identified using convolutional neural networks (CNNs) and the Mel Frequency Cepstral Coefficient (MFCC) approach for extracting characteristics from voice. Finally, the simulations showed that the proposed MFCC-CNN outperformed existing models in terms of performance.

Index terms: CNN, speech emotion recognition, facial emotion, speech emotion

1. INTRODUCTION

The automated recognition of emotions through the use of facial expressions involves three stages: face recognition, feature extraction and classification, and voice sound and hand gestures. Nevertheless, the most recent developments in human user interfaces, which have gone beyond the standard mouse and keyboard to automated speech recognition technologies and specialized interfaces designed for people with disabilities, do not fully account for these essential interactive capabilities, occasionally leading to less-than-ideal experiences. Robots could be able to assist humans accurately and effectively in ways that are more in line with their tastes and expectations if they could understand these emotional cues. Six archetypal emotions—shock, fear, disgust, fury, joy, and sadness—can be used to classify a variety of human emotions, according to psychological studies. Certain emotions are most effectively communicated through voice tone and facial expression. The study of emotions has become a crucial area of study that may help with a variety of goals by offering some insightful data. Whether intentionally or unintentionally, people express their emotions via their words and facial expressions. Several different types of knowledge, including voice, writing, and visual, can be used to interpret emotions. While olden times, speech and facial expression are important for useful tool for identifying emotions and have revealed many aspects, including mentality. It's a tremendous and difficult endeavour to uncover the emotions concealed underneath these

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

words and facial expressions. Researchers from a range of disciplines are striving to create techniques for more precisely understanding human emotions from multiple ways including speech and facial expressions, to meet this difficulty. Computer intelligence, natural language modelling systems, and additional methods are used to improve the reaction to various speeches and vocal-based techniques.

Numerous specific circumstances may benefit from the analysis of feelings. One such area is collaboration with actual computers. Customers may use computers to identify emotions, make better decisions, and increase the realism of interactions between humans and robots. We will look at the characteristics, constraints, and prospective directions of the current emotion detection methods, emotion models, and emotion databases in this work. We focus on evaluating labour jobs that need speech and face recognition in order to gauge emotions. We looked at the many technical sets that make up contemporary methods and technologies. The industry's key accomplishments have been made, and workable alternatives for better outcomes have been found.

2. LITERATURESURVEY

Due to the fast improvement of artificial intelligence technology, interest in FER research has significantly increased over the past few decades. For FER systems, many feature-based techniques are researched. Present techniques locate a facial region in a photo and take physical or geometric cues from it. One of the geometric characteristics is frequently the link between the various facial components. Facial landmarks can serve as instance of geometric properties [2, 30, 31]. The universal aspects of the face areas or other kinds of data about the facial regions are retrieved as emergence characteristics [20, 36]. The universal futures commonly exhibit PCA, a local binary pattern histogram, and other characteristics. Other revision separated the face district into distinct local regions and retrieved region-particular emergence traits [6, 9]. The key locations inside these local zones are first identified, which increases the identification accuracy. Due to the rapid advancement of deep-learning techniques, the CNN and recurrent neural network (RNN) are used in a variety of computer vision applications recently. In particular, CNN has performed exceptionally well in a variety of studies, including FER [10, 16, 44], face recognition, and object identification. Even though deep-learning-based algorithms have outperformed traditional approaches, challenges like micro-expressions, temporal fluctuations of expressions, and others are still difficult to overcome [21]. One of the most organic ways that humans may communicate is through speech signals, and they can be quickly and easily evaluated. Language-specific information and implicit paralinguistic information, such as speaker emotion, are both present in speech signals. because end-to-end learning (i.e., one-dimensional CNNs) are unable to do this mechanically for beneficial characteristics as acoustic characteristics can, most speech-emotion comprehension techniques, unlike FER, obtain Acoustic characteristics. Therefore, it's crucial to combine the appropriate audio features. It has been demonstrated in several research [1, 5, 14, 18, 27, 32, 34] that there is a connection between acoustic qualities and emotional voices. Speech-based emotion identification has a lower rate of recognition than other emotion-recognition methods, such face recognition, since there isn't a clear and predictable mapping among the emotional state and audio components. Choosing the optimum set of characteristics is therefore an important step in speech-emotion detection. Computers are capable of correctly and realistically identifying human emotions from voice broadcasts and facial photos. This necessitates varying degrees of suitable emotion data fusion. Three technologies—feature combination, judgement fusion, and model concatenation—are at the heart of most multimodal investigations. Incorporating various inputs can be made possible by deep-learning technology, which is used in many industries [7, 22]. A simple method for merging models with diverse inputs is model concatenation. Individually encoded tensors are produced by models with different data inputs. The combine function can be used to combine the tensors for every type. Speech signals were transformed by Yaxiong et al. into mel-spectrogram pictures so that the images were able to be used as input by a 2D CNN. Additionally, they posted the picture of the expression to a 3D CNN. They used a network of deep beliefs to combine multimodal emotion data in a very nonlinear way after combining the

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

two networks [28]. Analysis of the grouping that each model produces and re-identification utilising the particular criteria are the goals of decision fusion. This is accomplished through combining the SoftMax functionalities of the different kinds of networks and computing the dot product utilising weights, wherein the sum of the weights is 1. As a way to conduct speech-emotion detection, Xusheng et al. created a bimodal fusion technique wherein both voice data and face emotions are correctly merged. They merged the CNN and RNN models to translate speech inputs into features. Speech and facial expression data were combined using the weighted-decision fusion approach [40]. In order to collect simultaneous temporal face characteristics and temporal geometry data, Jung et al. employed the deep temporal appearances network with the deep temporal geometry networks [17]. They integrated both of these networks of distinct features and added the last layers of the fully linked portion of the networks before pre-training the networks in order to increase the performance of their model. A more complete fusion model has to be created because the bulk of current approaches simply include surface fusion [28].

3. METHODOLOGY

Emotion is a crucial aspect of human communication. Our actions and decision-making are influenced by a combination of our emotions, behavior, and thoughts since these three things are intertwined. This has led to a rise in interest in this area of science during the past several years. To improve them, automatic emotion recognition may be used in a variety of contexts. Consider human-computer interaction, where determining the customer's exciting state will enable the creation of a more organic, fruitful, and intelligent connection. Monitoring of human-human interactions is another topic since it allows for the early detection of disputes or undesirable circumstances. The automated emotion identification from voice, faces, and movies is also covered in this study. The suggested methodology made use of deep learning CNN techniques such corpus development, feature selection, proper classification scheme design, and information fusion with other sources of data like text.

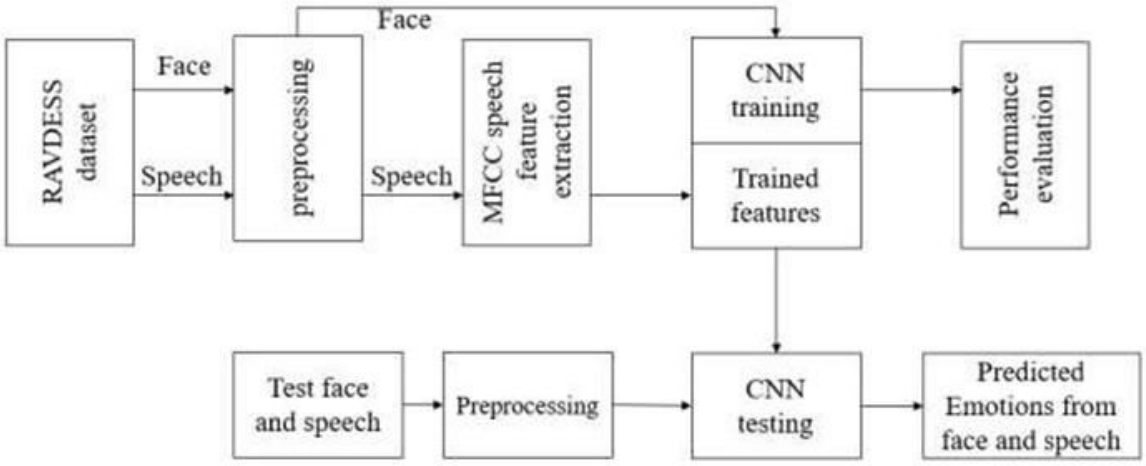

Fig. 1. Proposed block diagram

The suggested block design for emotion identification using voice and faces is shown in Figure 1. The RACVDESS dataset, which includes voice and facial data files, is taken into consideration for use in this work. The sounds from the voice files and facial photos were then eliminated during the pre-processing step on both datasets. Then, only speech data are used to extract MFCC features. The CNN model is then trained using pre-processed face data as well as speech-based MFCC characteristics. Finally, test features are applied to test face and speech data and contrasted by means of the pre-trained CNN model skin texture. lastly, using both facial and speech data, our AI-CNN model determines the expected emotion.

3.1 IMAGE AND SPEECHPRE-PROCESSING

Using computer algorithms to do image processing on digital pictures is known as digital image processing.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

Digital image processing, a branch of digital signal processing, provides significant settlement over analogue image processing. It enables the application of a considerably larger variety of algorithms to the supplied data. In order to provide our AI-Computer Vision models better data to work with, digital image processing aims to progress the picture data (features) by contains undesirable alteration and/or enhancing some key image properties. Our photos must be the same size as the network's input for the network to be trained and to make predictions on fresh data. We can rescale or scale down emotion data to the necessary size if we need to change the size of the photos to fit the network. By applying randomized augmentation to the data, we may successfully amplify the quantity of training data. Additionally, amplification makes it possible to train networks to be resistant to image data distortions. For instance, we may add random rotations to the input photos to make the network insensitive to the rotational state of the input images. A simple method for applying a small number of augmentations to 2-D pictures for classification issues is to use an augmented image data store. Image data can be kept in a table, an Image Data store object, or a numeric array. An image largest selection can be utilized for importing information in batches from photo collection which are too large to fit in memory. You can use a scaled 4-D array for training, forecasting, or categorization, or an improved image data store. 3-D arrays that have been shrunk are only useful for categorization and forecasting. Image data may be adjusted in one of two ways to fit the input size of a network. An image's height and width are increased by a scaling factor whenever it is enlarged. Expanding alters the spatial extends of the pixels in addition to their aspect ratio if the scaling factor in both the horizontal and vertical axes differs.

Cropping: Although removing a portion of the image, it keeps the dimension of each pixel. Images may be trimmed from any location, even the middle. A image is composed of a two-dimensional array of numbers (or pixels) having values between 0 and 255. The mathematical function f(x,y), wherein x and y are the two coordinates for the horizontal and vertical axes, accordingly, provides a definition for it.

Resize image: so as to visualize the alter, we will write two functions to show the photos in this phase, the first of which will exhibit one image and the second of which will display two images. Then, processing is a function created that only accepts the images as a parameter. Since some of the images captured by cameras and used in the pre-processing stage need to be resized due to size variations, we ought to found a base size for every photos fed into our AI techniques.

MFCC feature extraction

The earliest step of extraction is pre-emphasis. It is the process of increasing high frequency energy. The reason behind this is that lower frequencies in the vocal spectrum have more energy than higher frequencies. The nature of the glottal pulse is what causes this phenomenon, known as spectral tilt.

1) The Acoustic Classifier receives additional data as high-frequency energy increases, therefore improves phone detection ability. MFCC can be extracted by utilizing the method detailed following.
2) The provided speech signal is divided into frames that last for around 20 milliseconds. Typical frame intervals range from 5 to 10 milliseconds.
3) The previously mentioned frames are multiplied employing a Hamming window to guarantee the signal's continuation. The Gibbs effect is avoided by the usage of hamming windows. To ensure continuity at every frame's beginning and conclusion and avoid rash alterations at the end point, the Hamming window is multiplied to each frame of the signal. Furthermore, each frame has a hamming window that groups adjacent frequency components of like frequencies.
4) 4) A Mel-scale filter bank is put in place to the DFT power spectrum to produce the Mel spectrum. Mel-filter focuses greater on the critical area of the spectrum in order to obtain data values. Triangular band pass filters, like those in the human listening system, make form the Mel-filter bank. overlapping filters make form the filter bank. Every result produced by a filter is composed

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

of the total energy of several frequency bands. This method simulates the higher sensitivity to lower frequencies of the human ear. Getting the frame's vitality is a further essential aspect. Find the logarithm of the Mel-filter bank's rectangular output. the way that people react to log scales of signal strength. High levels of energy make people less susceptible to modest energy shifts than when energy levels are low. Logarithm compresses dynamic range of values.

5) 5) Mel-scaling and smoothness (pull to right). The mel scale is essentially linear under 1 kHz and exponential above 1 kHz.
6) 6) DCT is a further stage in the MFCC process that transforms the signal between the frequency domain onto the time domain and reduces data redundancy, so it can overlook the signal's minuscule temporal fluctuations. Using DCT to the mel-spectrum's logarithmic results in the mel-cepstrum. Minimise the amount of feature dimensions by using DCT. It lessens the spectral relationship among the coefficients of the filter banks. Low complexity and the 17 uncorrelated characteristics are benefits for any quantitative classifier. In the cepstral coefficients, no energy is gathered. As a result, adding an energy function is essential. 12 Mel Frequency Cepstral and one (1) energy coefficient are extracted as a consequence. The above-mentioned thirteen (13) facial appearance are taken to as the "basic features."
7) Obtain MFCC features.

Using the equation, the MFCC is created by transforming the frequency into the cepstral coefficients and the cepstral coefficients back into the frequency..

$$mel(f) = 2595 \times \log 10\ (1 + \ ^{f}) \Big|_{700}$$

Here f stands for the frequency in Hz that was used in the MFCC computation step. The formula that follows is used for calculating the MFCC characteristics..

$$C_n = \sum_{n=1}^{K} (logS_k) \lceil n(K -$$

where$n$=1,2,…..K

The DCT in this case does not include C0 since it reflects the mean value of the input speech signal, which is devoid of meaningful speech-related information. K indicates how many Mel cepstral coefficients there are. Each of the 20 ms worth of overlapping voice frames results in the creation of an audio vector made up of MFCC. This set of coefficients both represents and acknowledges the characteristics of speech..

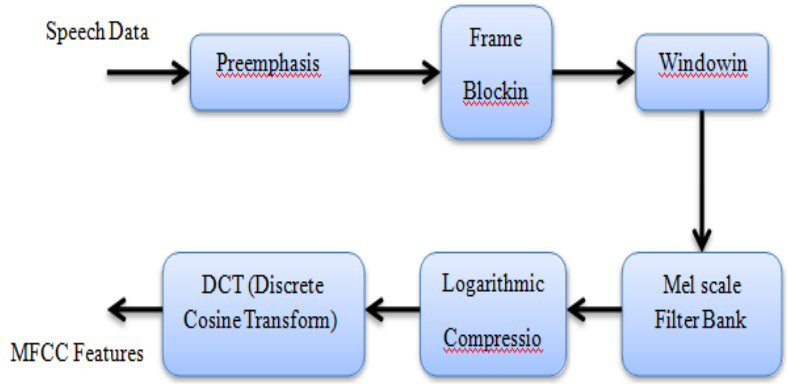

Fig. 2. MFCC operation diagram

CNN model

The deep CNN model for speech recognition-based emotion identification is shown in Fig 3; the suggested

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

deep CNN representation for emotion exposure using facial expressions is revealed in Fig 4.

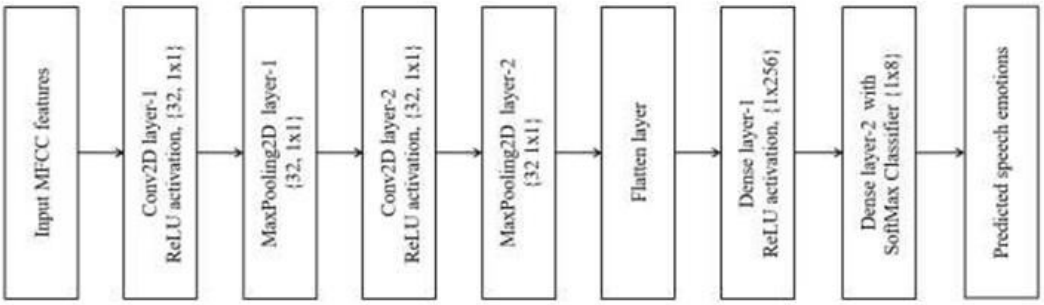

Fig.3.Deep CNN representation is suggested for voice recognition-based feeling detection.

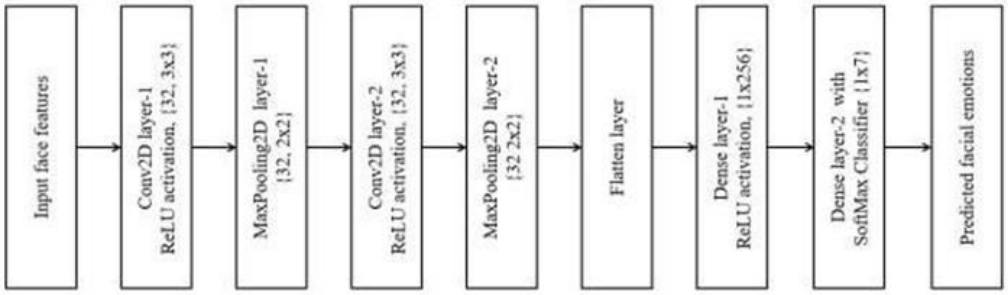

Fig.4.Projected deep CNN representation for facial expression-based feeling recognition

4. SIMULATIONRESULTS

Dataset

We utilised 28,709 photos with 7 distinct face emotion recognition models, including angry, happy, neutral, sad, disgusted, terrified, and startled. The voice emotion identification model employs the Ryerson Audio-Visual Database of exciting voice and Song (RAVDESS) dataset. Speech audio-only files from the RAVDESS have a data rate, sampling frequency, and arrangement of 16bit, 48kHz, and.wav. There are 1440 files in this section of the RAVDESS. 24 actors times 60 trials each equals 1440. The RAVDESS has 24 trained actors—12 male and 12 female—who each speak two lexically related sentences with a neutral North American accent. Speech expressions can be composed of expressions of calmness, joy, sadness, anger, fear, surprise, and disgust. here are two exciting intensity levels (normal and strong) and one neutral appearance created for each expression.

4.1 PERFORMANCE

Fig Figure 5 displays the example test photos for classifying feelings using given expressions. The simulated images depict every emotion possible, such as sadness, fury, contempt, surprise, and panic. Figure 6 shows the suggested deep CNN's prediction precision and decreased performance utilising facial expression, audio, and videos. The suggested deep CNN outperformed both facial expression and verbal inputs for emotion prediction using videos, as been observed from the two photos.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

Fig.5.Examples of emotion forecasting test pictures.

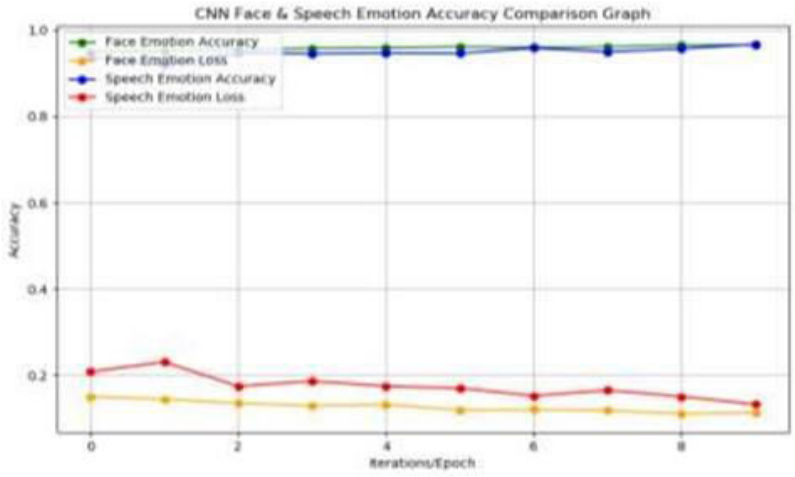

Fig.6. Accuracy and loss comparison of proposed CNN with speech and facial expression.

CONCLUSION

Whether done consciously or not, people convey their feelings via their speech and facial expressions, making the study of emotions an important field of research that may aid in a number of objectives. Voice, writing, and pictures are only a few of the informative mediums that may be used to understand emotions. In order to improve prediction accuracy and decrease loss, this article developed a deep CNN representation for emotion forecast as of speech and facial expression. Additionally, MFCC was worn to remove features from the provided speech samples for the speech CNN model.

REFERENCES

1. Bjorn S, Stefan S, Anton B, Alessandro V, Klaus S, Fabien R, Mohamed C, Felix W, Florian E,ErikM,MarcelloM,HuguesS,AnnaP,FabioV,SamuelK(2013)Interspeech2013ComputationalParalinguisticsChallenge: Social Signals, Conflict, Emotion, Autism
2. DeepakG, JoonwhoanL(2013) Geometric feature- based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensors 13:7714–7734.
3. Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos J, Delahoz EJ, Contreras-OrtizS(2020)Amachine learning model for emotion recognition from physiological signals. Biomed Signal Proces 55:101646
4. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and data bases. Pattern Recogn 44:572–587.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

5. Eyben F, Scherer KR, SchullerBW et al (2016) The Geneva minimalistic acoustic parameterset (geMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7:190–202.
6. Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local regionspecificfeatures andsupportvector machines.MultimedToolsAppl76:7803–7821.
7. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MITPress.https://www.deeplearningbook.org. Accessed 1Mar2020
8. Hamm J, Kohler CG, Gur RC, Verma R (2011) Automated facial action coding system fordynamicanalysisoffacialexpressionsinneuropsychiatricdisorders.JNeurosciMethods200:237–256
9. Happy SL, George A, Routray A (2012) A real time facial expression classification systemusinglocalbinarypatterns.InProc 4th Int Conf IntellHuman ComputInteract27–29:1–5
10. HasaniB,MahoorMH(2017)Facialexpressionrecognitionusingenhanceddeep3Dconvolutionalneuraln etworks.IEEEConfComputVisionPatternRecognitWorkshops(CVPRW).
11. HeJ,LiD,BoS,YuL(2019)Facialactionunitdetectionwithmultilayerfusedmulti-taskandmulti-labeldeep learningnetwork. KSIITransInternetInf Syst7:5546–5559.
12. HossainMS,MuhammadG(2019)Emotionrecognitionusingdeeplearningapproachfromaudio–visual emotionalbigdata. Inf Fusion 49:69–78.
13. HuttoCJ,EricG (2014)VADER:Aparsimonious rule-basedmodelforsentimentanalysis ofsocial mediatext. AAAIPublications, EighthIntAAAIConf WeblogsSocMedia
14. IliouT,AnagnostopoulosC-N(2009)Statisticalevaluationofspeechfeaturesforemotionrecognition.In: DigitaltelecommunicationsICDT'094thIntConfIEEE 121–126
15. JiaX,LiW,WangY,HongS,SuX(2020)Anactionunitco-occurrenceconstraint3DCNNbasedactionunitrecognition approach.KSIITrans InternetInfSyst14:924–942.
16. JosephR,SantoshD,RossG,AliF(2015)YouOnlyLook Once:Unified,Real-TimeObjectDetectionarXivpreprintarXiv:1506.02640
17. JungH,LeeS,YimJ,ParkS,KimJ(2015)Jointfine-tuningindeepneuralnetworksforfacialexpressionrecognition.2015 IEEEInt ConfComput Vision(ICCV).
18. Kao YH, Lee LS (2006) Feature analysis for emotion recognition from Mandarin speechconsideringthespecialcharacteristicsofChineselanguage.In: InterSpeech
19. Kaulard K, Cunningham DW, Bülthoff HH, Wallraven C (2012) The MPI facial expressiondatabase—A validated database of emotional and conversational facial expressions. PLoS One7:e32321.
20. Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facialexpressionrecognitionforlowresolutionimages.PatternRecognLett 34:1159–1168.
21. Ko BC (2018) A brief review of facial emotion recognition based on visual information.Sensors 18.
22. LeCunY,BengioY,HintonG(2015)Deeplearning,Nature521.
23. Lee C, Lui S, So C (2014) Visualization of time-varying joint development of pitch anddynamicsfor speech emotion recognition.JAcoustSoc Am135:2422.
24. Li S, Deng W (2020) Deep facial expression recognition: A survey. IEEE Trans AffectiveComp(EarlyAccess).
25. Liu M, Li S, Shan S, Wang R, and Chen X (2014) Deeply learning deformable facial actionparts model for dynamic expression analysis. 2014 Asian Conference on Computer Vision(ACCV)143–157.
26. LotfianR,BussoC(2019)Curriculumlearningforspeechemotionrecognitionfromcrowdsourced labels. IEEE/ACM TransAudio,SpeechLangProcessing4.
27. LuengoI,NavasE,HernáezI,SánchezJ(2005)Automaticemotionrecognitionusingprosodicparameters. In:Interspeech,493–496

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, Dec 2021

28. MaY,HaoY,ChenM,ChenJ,LuP,KoširA(2019)Audio-visualemotionfusion(AVEF):Adeep efficientweighted approach. Inf Fusion46:184–192.
29. MehrabianA(1968)Communicationwithoutwords.PsycholToday2:53–56
30. Mira J, ByoungChul K, JaeYeal N (2016) Facial landmark detection based on an ensemble oflocal weightedregressorsduringrealdrivingsituation.Int Conf PatternRecognit1–6.
31. Mira J, ByoungChul K, Sooyeong K, JaeYeal N (2018) Driver facial landmark detection inreal drivingsituations.IEEE Trans CircuitsSystVideo Technol 28:2753–2767.
32. Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using globalandlocalprosodic features.IntJSpeechTechnol16(2):143–160
33. SchererKR(2003)Vocalcommunicationofemotion:Areviewofresearchparadigms.Speech Comm40:227–256.
34. Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect inspeech:stateoftheartandlessonslearntfromthefirstchallenge.SpeechComm53(9–10):1062–1087.
35. Shaqr FA, Duwairi R, Al-Ayyou M (2019) Recognizing emotion from speech based on ageandgenderusinghierarchicalmodels. ProcediaComput. Sci. 151:37–44.
36. Siddiqi MH, Ali R, Khan AM, Park YT, Lee S (2015) Human facial expression recognitionusing stepwise linear discriminant analysis and hidden conditional random fields. IEEE TransImage Proc 24:1386–1398.
37. SongP,ZhengW(2018)Featureselection-basedtransfersubspacelearningforspeechemotionrecognition. IEEETrans. Affective Comput.(EarlyAccess)
38. Sun N, Qi L, Huan R, Liu J, Han G (2019) Deep spatial-temporal feature fusion for facialexpressionrecognition instatic images. Pattern RecognLett119:49–61.
39. Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speechemotionrecognition:Areview. Int JSpeechTechnol21:93–120.
40. Wang X, Chen X, Cao C (2020) Human emotion recognition by optimally fusing facialexpressionand speechfeature.SignalProcess ImageCommun.
41. WuCH,YehJF,ChuangZJ(2009)Emotionperceptionandrecognitionfromspeech,AffectiveInf Processing93–110.
42. Xiong X and Fernando DlT (2013) Supervised descent method and its applications to facealignment.2013 IEEE ConfComputVisionand PatternRecognit(CVPR).
43. Zamil AAA, Hasan S, Baki SJ, Adam J, Zaman I (2019) Emotion detection from speechsignals using voting mechanism on classified frames. 2019 Int Conf Robotics, Electr SignalProcessingTechnol(ICREST).
44. Zhang H, Huang B, Tian G (2020) Facial expression recognition based on deep convolutionlong short-term memory networks of double-channel weighted mixture. Pattern Recogn Lett131:128–134.
45. ZhangS,ZhangS,HuangT,GaoW(2008)Speechemotionrecognitionusingdeepconvolutionalneuralnetw orkanddiscriminanttemporalpyramidmatching.IEEETrans.Multimed. 20:1576–1590.
46. Zhang T, Zheng W, Cui Z, Zong Y, Yan J, Yan K (2016) A deep neural network-drivenfeature learning method for multi-view facial expression recognition. IEEE Trans. Multimed.18:2528–2536.
47. Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1D & 2D CNN LST Mnetworks.Biomed SignalProcessingControl47:312–323.