International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-4, April 2021

Bimodal Emotion Recognition using Machine Learning

Manisha S, Nafisa Saida H, Nandita Gopal, Roshni P Anand

Abstract*: The predominant communication channel to convey relevant and high impact information is the emotions that is embedded on our communications. Researchers have tried to exploit these emotions in recent years for human robot interactions (HRI) and human computer interactions (HCI). Emotion recognition through speech or through facial expression is termed as single mode emotion recognition. The rate of accuracy of these single mode emotion recognitions are improved using the proposed bimodal method by combining the modalities of speech and facing and recognition of emotions using a Convolutional Neural Network (CNN) model. In this paper, the proposed bimodal emotion recognition system, contains three major parts such as processing of audio, processing of video and fusion of data for detecting the emotion of a person. The fusion of visual information and audio data obtained from two different channels enhances the emotion recognition rate by providing the complementary data. The proposed method aims to classify 7 basic emotions (anger, disgust, fear, happy, neutral, sad, surprise) from an input video. We take audio and image frame from the video input to predict the final emotion of a person. The dataset used is an audio-visual dataset uniquely suited for the study of multi-modal emotion expression and perception. Dataset used here is RAVDESS dataset which contains audio-visual dataset, visual dataset and audio dataset. For bimodal emotion detection the audio-visual dataset is used.

Keywords*: Emotion recognition, Bimodal analysis, machine learning, Ensemble learning, k-cross validation, MFCC.

I. INTRODUCTION

Emotions are a language independent means of communication universally that are expressed non verbally. The mental state of a person can be recognized by recognizing the basic emotions through speech or facial expressio. In order to improve the accuracy, we are using bimodal approach using CNN. We also present you the evaluation and comparison of other experimented models like Random forest, Decision tree, CNN for speech and VGG, CNN, Xception for face. We take audio and image frame from the video input to predict the final emotion of a person. Facial emotion recognition is based on Xception model, which has different architecture than normal CNN [5]. Speech emotion recognition system is based on CNN model [15]. It recognizes emotion using the feature extraction. MFCC (Mel Frequency Cepstral Coefficients) is used as input to the model. The predicted emotion from the two modality is combined using the user score-based weighting [16]. Information obtained from different modalities can be fused using an integration scheme in a multimodal system [17]. A weighted average ensemble method is used. The technique used to allow different models to contribute in prediction of information in accordance to their trust or an estimated performance is defined as the weighted average ensemble. The facial emotion and the speech emotion is integrated using a User-score-based weighting technique. The predicted emotion score is obtained by classifying the emotions individually from two modalities, and using the fixed score fusion rules including the sum, product and the weighted average, the score level fusion is performed [6]. The final score gives the predicted emotion from the bimodal system.

K.S Rao et al [8] proposed a system for recognizing emotions from video using facial and acoustic features. The acoustic features represent the spectral and prosodic information which is used for recognizing emotions from the speech signal. To capture the emotion-specific information from acoustic and facial features, an auto associative neural network models were used to capture. Five basic emotions were collected in a constrained studio environment and were considered for experimentation. However, the emotions are not restricted to these 5 emotions considered in this work in real life. K.M Aswin et al [9] proposed Human Emotion Recognition System (HERS), which is a real time system that analyses the facial expression and speech features for emotion recognition. The Gabor filter is used for image filtering and the emotions are classified using the Support Vector Machine (SVM). The Mel- Frequency Cepstral Coefficients (MFCC) vectors are computed from the real time voice input obtained from the microphone and computes. The MFCC vectors are given as the inputs to a linear SVM, which classifies it into 5 different emotions. The model provided a very good performance in detecting the emotions of “happiness”, “anger” and “surprise”. However, it was found to be less effective in detecting the emotions of “sadness” and “fear”. F.S Farahani et al [10] proposed a new fuzzy- based method for emotion recognition

Manuscript received on April 12, 2021.
Revised Manuscript received on April 19, 2021.
Manuscript published on April 30, 2021.
* Correspondence Author
Manisha S*, Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India. Email: manishas@ssn.edu.in
Nafisa H Saida, Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India.
Nandita Gopal, Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India.
Roshni P Anand, Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India.

from eyes and mouth features from data taken from people in different ages. This method detects the eyes and mouth region by combining the different color spaces. Eye opening, mouth opening, eye opening/width ratio and mouth width are the four parameters that were selected for fuzzy analysis.

K. Kushwah et al [11] proposed the automatic facial expression recognition system that includes a face detector, a facial feature extractor for mouth, left eye and the right eye, for facial characteristic extraction and facial expression recognition. A. Koduru et al [12] proposed a speech emotion recognition model that extracts the features and selects the required region of interest and classifies the emotion. The main focus of this work was to use different feature extraction algorithm to improve the speech emotion recognition rate. The MFCC, Discrete Wavelet Transform (DWT), pitch, energy and Zero crossing rate (ZCR) algorithms are used for extracting the features.

Y. Zhou et al [13] uses both spectral and prosodic features for recognizing the emotion through speech input. Both the spectral and the prosodic features contain emotion information, and combining of these spectral features and prosodic features will improve the performance of the emotion recognition system. The experimental results suggests that the combination of both spectral features and prosodic features fetches the emotion error reduction rate of 18 percent and 52.8 percent, over experiments that used only spectral and prosodic features. T.L. Nwe et al [14] proposed an emotion classification from speech using the text independent method. The proposed method uses the short time log frequency power coefficients (LFPC) to represent the speech signals and a discrete hidden Markov model (HMM) for classification of emotions. Performance of the LFPC feature parameters is compared with the linear prediction Cepstral coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) feature parameters are most commonly used features in speech recognition systems. Results suggests that the average accuracy of emotion classification is 78%

II. DATASET

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is used for training the model and performance of the model is evaluated using the test data from the same dataset. The RAVDESS dataset is a validated multimodal database that contains emotional speech and song as the data. The database contains 24 professional actors (i.e. 12 female, 12 male), and is gender biased. The actors vocalize two lexically matched statements in a neutral North American accent. The emotions of the speech include emotions such as neutral, happy, sad, angry, fearful, surprise, calm and disgust expressions, and the song data contains emotions such as calm, happy, sad, angry, and fearful emotions as shown in figure 2.

Normal and strong are the two emotional intensity at which each expression is produced. And Audio-only (AO), Audio-Video (AV), and Video-only (VO). For experimentation in this paper, the Audio-Video (AV) format are the formats in which the data is made available for each recorded production of an actor to recognize the emotions as shown in figure 1. These files contain 60 trials per actor x 1 modality (AV) x 24 actors = 1440 sets of data. Two distinct statements were vocalized by the actors in the speech conditions. The two statements ("Kids are talking by the door", "Dogs are sitting by the door") were spoken in eight emotional intentions. Statements were seven syllables in length and that matched both in word frequency and familiarity [1].

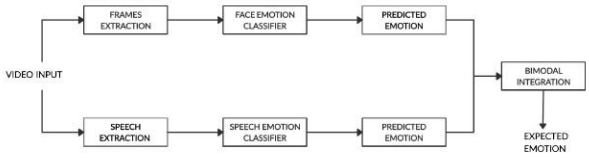

Fig. 1.Proposed Architecture for Bimodal Emotion Recognition

Fig. 2. Sample images from RAVDEES dataset

III. K-FOLD CROSS VALIDATION OF DATASET

The important performance metric to measure the classification of any data is accuracy. While working on small datasets, the ideal choices for validation and measuring of accuracy are k-fold cross-validation with large value of k (but smaller than number of instances) or leave-one-out cross-validation [2]. The cross- validation can be used to estimate the expected level of fit of a model to a data set which is independent of the data that were used to train the model. Estimation of any quantitative measure of fit that is appropriate for the data and model can be measured using this. In k-fold cross-validation, the original dataset is randomly partitioned into k equal sized samples resulting in a leave-one-out strategy [3]. Out of the k samples, a single sample is used as the validation data for testing the model, and the remaining k-1 samples are used as training data. The cross- validation process is then repeated k times, with each of the k samples used exactly once as the validation data. The average of these k results produced can be used as a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. In this paper, we take k=6 for facial emotion recognition model and for k=7 for speech emotion recognition model for the purpose of experimentation. The result obtained is that the single estimate for facial emotion recognition model is 57.02% after taking average of all k-fold cross-validation. Also the single estimate for speech emotion recognition model is 83.21% after taking average of all k-fold cross-validation.

IV. PROPOSED SYSTEM

A. Image and Audio extraction from Video

From the video dataset it is necessary to extract the key frame which contain the emotion as the video starts neutrally, then utters one of the statements with one of the emotions then ends with neutral emotion. Video is converted into frame images as shown in Figure 3.

Key frame is extracted in order to obtain the most suitable frame containing the emotion as shown in Figure 4. Image is converted from RGB to grayscale. Haar cascade is applied for face detection and to exclude the background of the image as shown in Figure 5. the image is resized to 48x48 pixel size as shown in Figure 6. Speech (the audio) is extracted from the video using the movie py editor as shown in Figure 7. The extracted audio and image are used for training the bimodal system. The bimodal emotion recognition system consists of three major parts: audio processing, visual processing and data fusion (audio, video) for detecting the emotion of a person as shown in Figure 8. Ravdess Dataset is used for training the model for both facial emotion detection as well as speech emotion detection.

Fig. 3.Proposed Architecture for Bimodal Emotion Recognition

Fig. 4. Key Frame Selection

B. Facial Emotion Recognition

For facial emotion recognition(image) we are using Convolution Neural Net- work with Xception model, that uses bounded face as input of image size 48 ∗ 48 pixels and predicts probabilities of 7 different emotions in the output layer [4]. The Xception model is special CNN architecture that has less overfitting problems than CNN models like VGG16[7]. We have also trained the images on VGG16 model and CNN model with different depths and max pooling layers [10], and concluded to use Xception model, as it has higher accuracy than VGG16 and CNN. Both the models are trained using the preprocessed images to predict the emotion in the output layer. VGG16, CNN and Xception model was trained using different batch size and epoch. The least accuracy was obtained for VGG16 model, 32.07%. The CNN model got accuracy almost close to Xception model but the "fear", "sad", and "surprise" was misclassified. Hence the Xception is used for facial emotion recognition. Xception model is created as a combination of two most successful experimental assumptions in CNNs: the depth wise convolutions and the use of residual modules. Xception model contains the following set of layers: 4 Conv2D layer,1 Global Avg Pooling 2D, 1 Softmax, center block contains 2 Separable Conv2D, 1 Conv2D, and 1 Maxpool2D.

This center block is repeated 4 times which separates Xception from another CNN model [4]. Conv2D – The convolution operation is performed in this 2D Convolutional layer. The layer creates a convolution kernel which convolves with the layer input and produces a tensor of outputs. The kernel size used in this experimentation is of size 3x3 and Rectified Linear Unit (ReLU) is used as the activation function.

– Batch Normalization- To normalize the input in a scale of o to 1 values, the batch normalization operation is performed on inputs that are given to the next layer to avoid the values scattered all over the place.

– MaxPooling2D – In this model built the function uses a pooling window of size 2x2 with 2x2 strides to perform the pooling operation on the data.

– Softmax – This function normalizes K real numbers taken from the input vector into a probability distribution consisting of K probabilities.

Fig. 5. RGB to Grayscale Conversion of Key Frame

Fig. 6. Key Frame Resized to 48*48

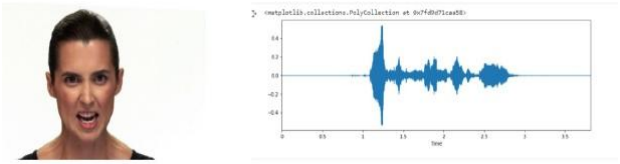

Fig. 7. Audio Extraction from Input Video

C. Speech Emotion Recognition

Speech Emotion Recognition We will train a CNN model with 7 layers- 6 Conv2D layer and 1 Dense layer. Audio is converted into Mel Frequency Cepstral Coefficients (MFCCs) and used as input feature [5]. The features are extracted by changing the speech waveform to a form of parametric representation at a relatively minimized data rate that can be used for subsequent processing and analysis. We have trained other models like Decision Tree, Random Forest apart from CNN. Random Forest is the simplest and widely used algorithm. It is used for classification and regression. The Random Forest based approach showed the least prediction rate. Experimental results show that accuracy obtained from the

Decision tree is slightly higher than Random forest. CNN utilizes a layer wise framework and the accuracy obtained is 86.80% which is the highest accuracy obtained among the models used for speech recognition. So, CNN is chosen for recognizing the emotion from the extracted MFCC features.

Algorithm 1: MFCC Algorithm

Result: Extraction of MFCC features

- MFCC are cepstral coefficients derived on a frequency scale centered on human auditory perception. In the computation of MFCC, the first thing is windowing the speech signal to split the speech signal into frames
- Since the high frequency formats process reduced amplitude compared to the low frequency formats, high frequencies are emphasized to obtain similar amplitude for all the formats
- After windowing, Fast Fourier Transform (FFT) is applied to find the power spectrum of each frame. Subsequently, the filter bank processing is carried out on the power spectrum, using melscale as shown in Figure 9.

The MFCC feature extracted from the speech is given to CNN model. We consider only 2.5s of the audio duration to extract the MFCC feature. The CNN model consists of 4 Conv1d layer with filter size 8 followed by one Flatten layer, Dense layer and Maxpooling 2D layer. The activation function used is 'ReLu' (Rectified Linear Unit). It allows the network to converge very quickly.

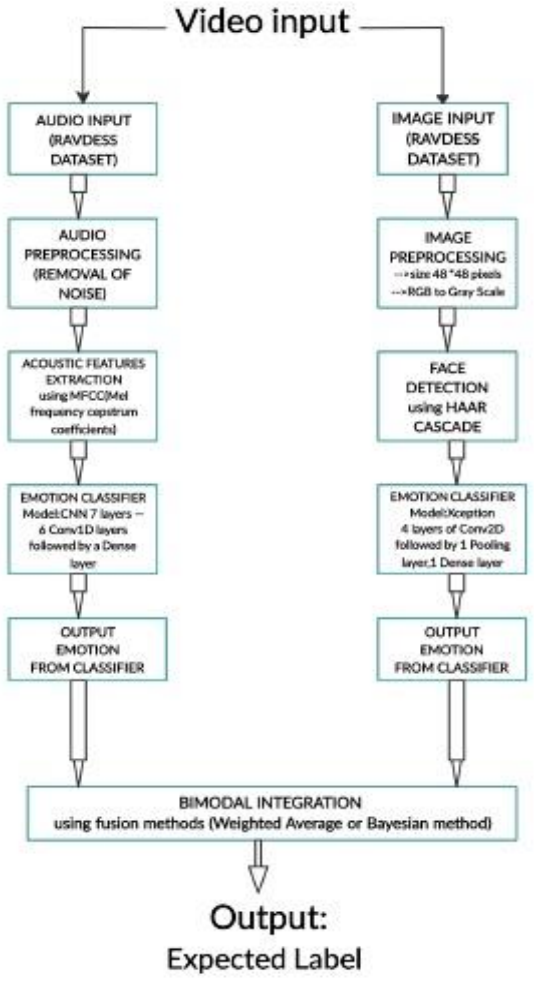

Fig. 8. Proposed System

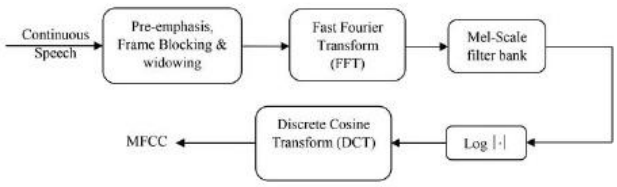

Fig. 9. MFCC Feature Extraction from Input Speech

D. Bimodal Integration using Fusion Rule

The two modalities are combined using a weighted average ensemble. Multiple models that contribute in predicting a data based on a proportion of the trust and their estimated performance can be done using a weighted average ensemble approach. In this approach of model averaging, an ensemble learning is achieved where each member of the ensemble model contributes equally to the final prediction of data. In this paper, the integration of the facial emotion model and the speech emotion model is done using a user-score-based weighting technique. The model weights used are small positive values such that the sum of all weights equals one. The values represent the percentage of trust or the expected performance measure for each model and the wights are used to represent the same. The predicted emotion score is obtained from two modalities, after individually classifying and the score level fusion is performed using three fixed score fusion rules (sum, product and Bayesian) [6]. The following are the formula used: Sc1, Sc2 are the score of image classifier and speech classifier respectively.

Sum: $S = Sc_1 + Sc_2$ (1)

Product $P = Sc_1 * Sc_2$ (2)

Bayesian :

$Sc_1 * Sc_2/[(1 - Sc_1) * (1 - Sc_2) + (Sc_1 * Sc_2)]$ (3)

Weighted sum: $s = wt * Sc_1 + (1 - wt) * Sc_2$. (4)

Weighted product: $s = Sc_1^{wt} * Sc_2^{1-wt}$. (5)

weight wt was computed by maximizing the system performance on a given operational point.

Weighted Average:

$s = wt_1 * Sc_1 + (1 - wt_1) * Sc_2$/No. of elements in array (6)

Out of all the methods, we are using the Weighted Average method (equation 6) and Bayesian method (equation 3) to integrate the predicted emotion to obtain the final emotion. The weights are equally We also analyze the results of the above two methods and conclude that Bayesian method is more inclined to higher accuracy model, whereas Weighted Average is more dependent on the assigned weights.

Table 1: Accuracies of Xception model on image testing set for emotion classification

ALGORITHM	BATCH SIZE	EPOCHS	ACCURACY
Xception	64	50	60.7%
Xception	64	150	68.63%

Table 2: Accuracies of CNN on image testing set for emotion classification

ALGORITHM	BATCH SIZE	EPOCHS	ACCURACY
CNN	16	150	52.02%
CNN	16	200	57.21%
CNN	32	150	63.04%
CNN	32	200	65.1%

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 3: Accuracy of Vgg on image testing set for emotion classification

ALGORITHM	BATCH SIZE	EPOCHS	ACCURACY
Vgg	64	200	32.07%

V. RESULTS AND DISCUSSIONS

Several experiments were performed to make most of the proposed system. The experiments were performed on several Machine Learning models like VGG16, CNN, Xception for facial emotion recognition using different epochs and batch size, then the models were evaluated and compared as shown in tables 1, 2 and 3. The model with highest accuracy was considered as the final model for face emotion detection. The following tables gives the accuracy of different models of facial emotion recognition. Comparisons are made among different epoch of the Xception model. Also, comparisons are made among different epoch and batch size of the CNN model. The following results are obtained:

– The accuracy obtained is slightly lower than the Xception model.

– Vgg model was also utilized but the accuracy obtained was the least among all the trails.

– Vgg works for large datasets, as the number of images in our dataset is less, low accuracy was obtained.

Out of all the models for recognizing the emotions from facial features, the Xception model is chosen. Similarly different models were also experimented for Speech emotion recognition. The models like Decision trees, Random Forest apart from CNN were evaluated and compared by their accuracy as shown in table 4. The model with highest accuracy was considered as the final model for speech emotion detection. The following table gives the accuracy of different models of speech emotion recognition.Face Emotion Recognition (video frame) yielded an accuracy of 68.63%, whereas Speech Emotion Recognition yielded an accuracy of 86.80%. The Bimodal Emotion Recognition system yielded an accuracy of 73.17%. The Face Emotion Recognition model developed using K-fold Validation obtained an accuracy of 57.02% and the Speech Emotion Recognition model developed using K-fold Validation obtained an accuracy of 83.21%.

Table 4: Accuracy comparison amongst different models on speech testing set for emotion classification

ALGORITHM	ACCURACY
Random Forest	58.8%
Decision Tree	61.11%
CNN	86.8%

The predicted emotion from each modality, from facial emotion recognition and speech emotion recognition is plotted as graph. After bimodal integration the final emotion is predicted and plotted as graph as shown in Figure 10 and 11.

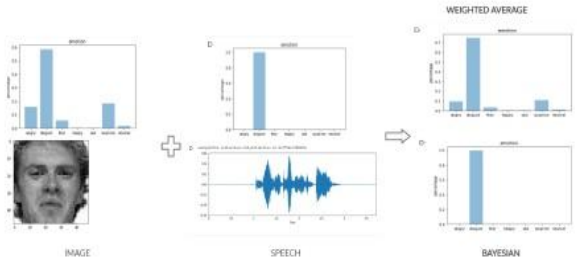

Fig. 10. Predicted Output for Disgust Emotion

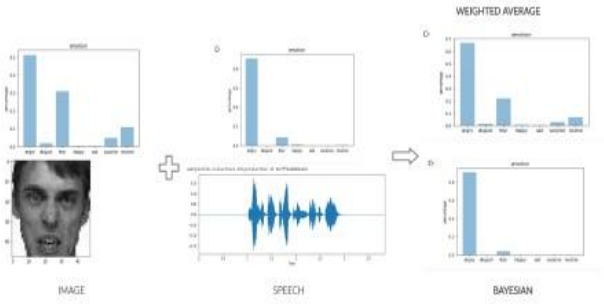

Fig. 11. Predicted Output for Angry Emotion

VI. CONCLUSION

This paper presented a bimodal emotion recognition system that uses information from the channels audio and visual data obtained from a video stream. In this study, we have proposed a fusion model that has a speech emotion recognition system based on MFCC Transform. These features extracted were used in modelling the proposed CNN algorithm. Similarly, the facial emotion recognition system is proposed using the Xception algorithm. Xception model is chosen over other techniques such as CNN and VGG based on computational grounds. These two modalities from speech and video are combined using the score level fusion using weighted average and Bayesian method. It is observed that weighted average method performed well compared to Bayesian method. In future, the integration of two modalities can be performed using the feature level fusion. For improving the accuracy of the face emotion recognition, the Hidden Markov Model (HMM) can be used. Each person's expression varies with respect to their current mental emotion. Also, some expressions such as sadness, fear and disgust can be ambiguous. Therefore, differentiating such ambiguous expressions becomes really hard. This information can be used as a major clue to improve the performance of emotion recognition in the future research.

REFERENCES

1. Livingstone, S.R. and Russo, F.A, 2018, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", PloS one, 13(5).
2. Yadav, S. and Shukla, S, 2016, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification", 2016, IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, pp. 78-83.

3. Pinto, A., Pereira, S., Dinis, H., Silva, C. A. and Rasteiro, D. M. L. D., 2015, "Random decision forests for automatic brain tumor segmentation on multimodal MRI images", 2015 IEEE 4th Portuguese Meeting on Bioengineering (ENBENG), Porto, pp. 1-5.
4. Arriaga, O., Valdenegro-Toro, M. and Ploger, P., 2017. "Real-time convolutional neural networks for emotion and gender classification,", arXiv preprint arXiv: 1710.07557.
5. Wang, K., An, N., Li, B.N., Zhang, Y. and Li, L., 2015, "Speech emotion recognition using Fourier parameters", IEEE Transactions on affective computing, 6(1), pp.69-75
6. Akhtar, Z., Fumera, G., Marcialis, G.L. and Roli, F., 2012, March, "Evaluation of multimodal biometric score fusion rules under spoof attacks", In 2012 5th IAPR International Conference on Biometrics (ICB) (pp. 402-407). IEEE.
7. Lo, W.W., Yang, X. and Wang, Y., 2019, June, "An Xception Convolutional Neural Network for Malware Classification with Transfer Learning", In 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS) (pp. 1-5). IEEE.
8. Rao, K.S. and Koolagudi, S.G., 2015, "Recognition of emotions from video using acoustic and facial features", Signal, Image and Video Processing, 9(5), pp.1029-1045.
9. Aswin, K.M., Vasudev, K., Shanty, K. and Sreekutty, I.K., 2016, August "HERS: Human emotion recognition system", 2016 International Conference on Information Science (ICIS) (pp. 176-179). IEEE.
10. Pathar, R., Adivarekar, A., Mishra, A. and Deshmukh, A., 2019, April. "Human Emotion Recognition using Convolutional Neural Network in Real Time", 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) (pp. 1-7). IEEE.
11. Kushwah, K., Sharma, V. and Singh, U., 2017, April. "Neural network method through facial expression recognition", 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 2, pp. 532-537). IEEE.
12. Koduru, A., Valiveti, H.B. and Budati, A.K., January 2020, "Feature extraction algorithms to improve the speech emotion recognition rate", International Journal of Speech Technology, pp.1-11.
13. Zhou, Y., Sun, Y., Zhang, J. and Yan, Y., 2009, December, "Speech emotion recognition using both spectral and prosodic features", 2009 International Conference on Information Engineering and Computer Science (pp. 1-4). IEEE.
14. Nwe, T.L., Foo, S.W. and De Silva, L.C., 2003, "Speech emotion recognition using hidden Markov models", Speech communication, 41(4), pp.603-623.
15. Zhang, B., Quan,C. and Ren,F.,2016,June, "Study on CNN in the recognition of emotion in audio and images", In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-5). IEEE.
16. Viriri, S., Tapamo, J. R. (2012), "Integrating iris and signature traits for personal authentication using user-specific weighting", Sensors, 12(4), 4324 - 4338.
17. Mukahar, N. B., Jaafar, H. B., Rosdi, B. A. B. (2017), "User Specific Weights Based on Score Distance and EER for Weighted Sum Rule Fusion", In 9th International Conference on Robotic, Vision, Signal Processing and Power Applications (pp. 185-191). Springer, Singapore.

AUTHORS PROFILE

Dr.. S. Manisha, is the Assistant Professor in the Department of Computer Science and Engineering in Sri Sivasubramaniya Nadar College of Engineering. She received her B.E. in Computer Science and Engineering degree from Anna University in the year 2010. She received her M.E (CSE) degree from Anna University in the year 2012. She received her Ph.D from Anna University in 2021. She is a member of ACM. Her research interests include Image Processing, Multimedia Steganography, Computer Vision, and Machine Learning.

Nafisa H Saida, completed her B.E. CSE from Sri Sivasubramaniya Nadar College of Engineering in 2021. Her research interests includes Image Processing and Deep Learning

Nandita Gopal, completed her B.E. CSE from Sri Sivasubramaniya Nadar College of Engineering in 2021. Her research interests include Image Processing and Deep Learning.

Roshni P Anand, completed her B.E. CSE from Sri Sivasubramaniya Nadar College of Engineering in 2021. Her research interests include Image Processing and Deep Learning.