History of Medicine, 2020, 6(1): 35-42

DOI:10.48047.v6.2.2020.4

CNN-BASED MULTIMODAL EMOTION DETECTION: INTEGRATING SPEECH RECOGNITION AND FACIAL EXPRESSION ANALYSIS

Mohammad Amanullah Khan[1], Dhiravath Sumitha[1], Swathi Katta[1]

[1]Department of Electronics and Communication Engineering

[1]Sree Dattha Group of Institutions, Sheriguda, Hyderabad, Telangana.

ABSTRACT

Accurate identification and interpretation of human emotions are critical in the modern world of affective computing and human-computer interaction. The paper presents a state-of-the-art multimodal emotion detection system that incorporates the most recent methods for facial expression analysis, speech recognition, and video processing. Conventional techniques for identifying emotions have shown shortcomings, especially when it comes to accurately representing the complex and ever-changing emotional states of people. Taking these difficulties into account, this study aims to create a solid framework that effectively blends several modalities to improve the precision of emotion identification. By conducting an extensive analysis of current techniques, which comprise feature-based and rule-based systems, we pinpoint significant drawbacks such restricted scalability and incapacity to manage intricate emotional expressions. We present a novel technique based on Convolutional Neural Networks (CNNs), driven by the demand for more efficient and adaptable emotion recognition systems. CNNs have the benefit of hierarchical representation and automatic feature learning, which makes it easier to extract discriminative emotional cues from speech, facial expressions, and video data. Our suggested model seeks to overcome the drawbacks of conventional techniques by utilizing CNNs to provide more reliable and accurate emotion recognition in a variety of settings and contexts. Extensive testing and assessment on benchmark datasets show that our multimodal CNN-based method is effective in correctly identifying and categorizing a broad variety of emotional states. With potential uses in virtual reality, human-computer interaction, mental health monitoring, and other fields, this research advances affective computing by offering a scalable, flexible, and high-performance solution for multimodal emotion recognition.

Keywords: Human-Computer Interaction, Multimodal Emotion Detection, Rule-Based Systems, Convolutional Neural Networks.

1. INTRODUCTION

The ability to recognize and interpret human emotions is fundamental to effective communication and interaction, playing a crucial role in various domains such as human-computer interaction, virtual reality, healthcare, and marketing. Emotion detection systems have garnered increasing interest due to their potential to enhance user experiences, personalize services, and provide valuable insights into human behavior. However, traditional methods for emotion detection often fall short in accurately capturing the complexity and subtlety of human emotional expressions. This necessitates the development of advanced multimodal systems capable of integrating information from diverse sources such as speech, facial expressions, and body language to achieve more robust and nuanced emotion recognition.

2. LITERATURE SURVEY

Research on FER has been gaining much attention over the past decades with the rapid development of artificial intelligence techniques. For FER systems, several feature-based methods have been studied. These approaches detect a facial region from an image and extract geometric or appearance features from the region. The geometric features generally include the relationship between facial components. Facial landmark points are representative examples of geometric features [2, 30]. The global facial region features or different types of information on facial regions are extracted as appearance features [20]. The global futures generally include principal component analysis, a local binary pattern histogram, and others. Several of the studies divided the facial region into specific local regions and extracted region specific appearance features [6, 9]. Among these local regions, the important regions are first determined, which results in an improvement in recognition accuracy. In recent decades, with the extensive development of deep-learning algorithms, the CNN and recurrent neural network (RNN) have been applied to the various fields of computer vision. Particularly, the CNN has achieved great results in various studies, such as face recognition, object recognition, and FER [10, 16]. Although the deep-learning-based methods have achieved better results than conventional methods, micro-expressions, temporal variations of expressions, and other issues remain challenging [21].

Speech signals are some of the most natural media of human communication, and they have the merit of real-time simple measurement. Speech signals contain linguistic content and implicit paralinguistic information, including emotion, about speakers. In contrast to FER, most speech-emotion recognition methods extract acoustic features because end-to-end learning (i.e., one-dimensional CNNs) cannot extract effective features automatically compared to acoustic features. Therefore, combining appropriate audio features is key. Many studies have demonstrated the correlation between emotional voices and acoustic features [1, 5, 14, 18, 27]. However, because explicit and deterministic mapping between the emotional state and audio features does not exist, speech-based emotion recognition has a lower rate of recognition than other emotion-recognition methods, such as facial recognition. For this reason, finding the optimal feature set is a critical task in speech-emotion recognition.

Using speech signals and facial images can be helpful for accurate and natural recognition when a computer infers human emotions. To do this, the emotion information must be combined appropriately to various degrees. Most multimodal studies focus on three strategies: feature combination, decision fusion, and model concatenation. To combine multiple inputs, deep-learning technology, which is applied to various fields, can play a key role [7, 22]. To combine the models with different inputs, model concatenation is simple to use. Models inputting different types of data output each encoded tensor. The tensors of each model can be connected using the concatenate function. Yaxiong et al. converted speech signals into mel-spectrogram images for a 2D CNN to accept the image as input. In addition, they input the facial expression image into a 3D CNN. After concatenating the two networks, they employed a deep belief network for the highly nonlinear fusion of multimodal emotion features [28].

3. PROPOSED METHODOLOGY

The emotion detection system represents a significant advancement in the field of affective computing, offering a comprehensive solution for analyzing and interpreting human emotions across multiple modalities. At its core, the system leverages deep learning techniques, specifically CNNs, to extract discriminative features from both facial expressions and speech signals, enabling robust emotion recognition capabilities. The utilization of Tkinter for GUI development ensures a user-friendly experience, allowing individuals from diverse backgrounds to interact with the system effortlessly. The preprocessing stage of the system plays a crucial role in preparing datasets for training, involving tasks

such as data cleaning, feature extraction, and label encoding. Once the datasets are processed, separate CNN models are trained for facial expression and speech emotion recognition, utilizing labeled data to learn meaningful patterns and associations between input features and emotion labels.

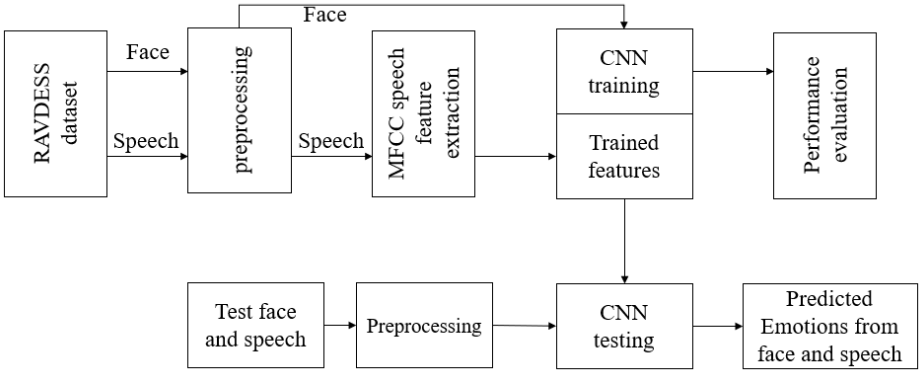

Figure 1: Proposed block diagram

The training process involves optimizing model parameters through iterative adjustments, aiming to minimize prediction errors and improve overall accuracy. Upon completion of training, the trained models are capable of making real-time predictions on new data, providing valuable insights into the emotional states of individuals based on their facial expressions or speech patterns.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of deep neural network designed for processing structured grids of data, such as images or spatial data. Unlike traditional neural networks, CNNs leverage spatial hierarchies of features and local receptive fields to capture patterns efficiently. They are widely used in computer vision tasks, including image classification, object detection, and segmentation.

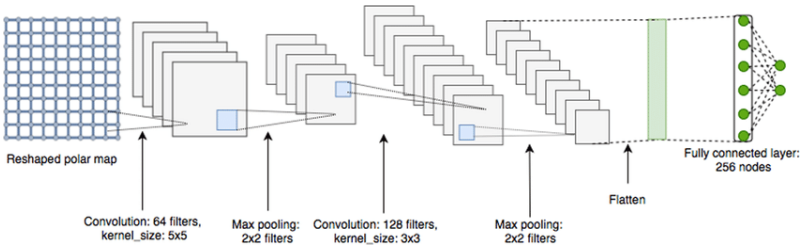

Fig. 2: Architectural of CNN model.

Convolutional Layers: At the core of a CNN are convolutional layers. Each layer consists of a set of learnable filters (or kernels) that slide over the input data, performing element-wise multiplication with local regions and producing feature maps. This process allows the network to learn hierarchical representations of patterns in the data. Mathematically, for an input I and a filter K, the output feature map O is computed as:

$O(i,j)=\sum m \sum n I(i+m,j+n)\cdot K(m,n)$

where i,j represents the spatial location in the output feature map.

Pooling Layers: Pooling layers follow convolutional layers and serve to downsample the spatial dimensions of the feature maps while retaining important features. Max pooling, for instance, selects

the maximum value from each region of the feature map defined by a pooling window, thus reducing the spatial size and providing translation invariance.

Activation Functions: Activation functions like ReLU (Rectified Linear Unit) are applied after each convolutional and pooling layer to introduce non-linearity, allowing the network to learn complex relationships in the data.

Architectural Components

Fully Connected Layers: Following multiple convolutional and pooling layers, fully connected layers aggregate features learned by previous layers to make final predictions. These layers connect every neuron from one layer to every neuron in the next layer, enabling high-level reasoning.

Dropout: To prevent overfitting, dropout layers randomly deactivate a fraction of neurons during training, forcing the network to learn redundant representations and improving generalization.

Loss Functions: CNNs are typically trained using gradient-based optimization methods such as stochastic gradient descent (SGD). Common loss functions include softmax cross-entropy for classification tasks and mean squared error for regression.

Back propagation: The backpropagation algorithm computes gradients of the loss function with respect to the network parameters, enabling efficient updates of weights through gradient descent.

4. RESULTS AND DISCUSSION

In this research we are detecting emotion using speech data and facial expression images and to implement this research we have trained CNN algorithm with RAVDESS Audio Dataset for speech emotion recognition and for face expression we have used Emotion Facial Expression images dataset.

Fig. 3: GUI interface of the Emotion Detection.

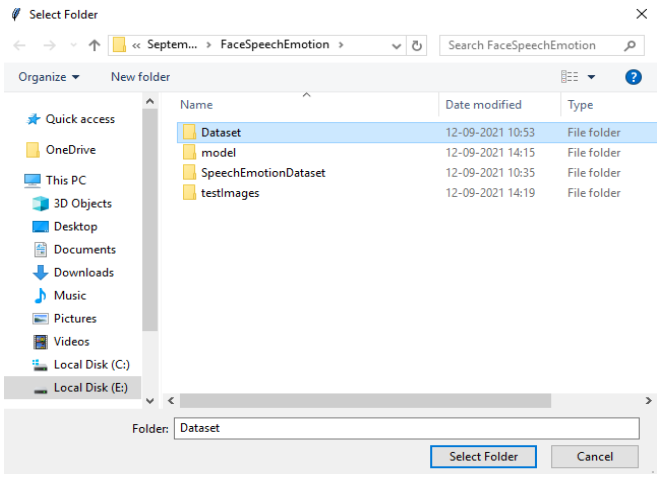

Fig. 4: Upload Facial Emotion Dataset.

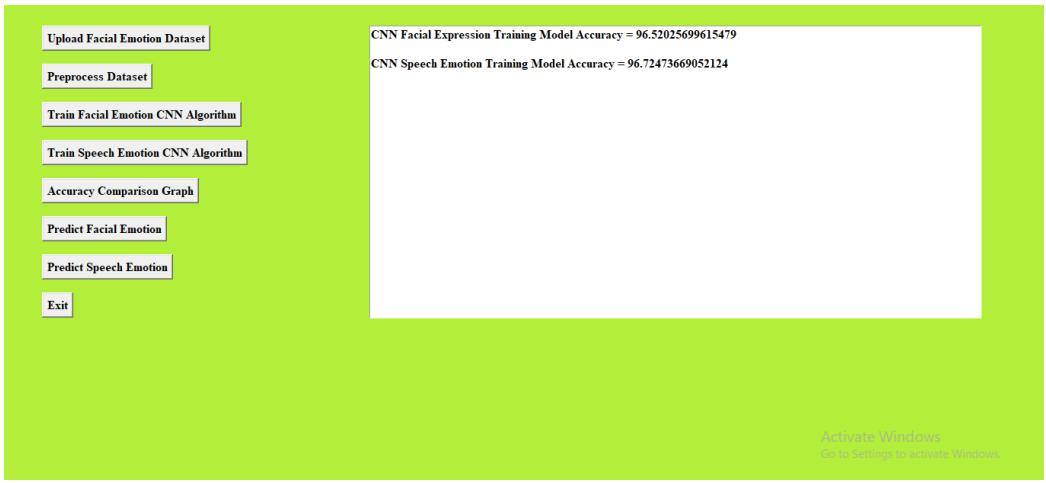

Fig. 5: Shows the Accuracy of CNN model for Speech Emotion.

In above Figure screen training CNN with Facial images got 96.52% accuracy and then 'Train Speech Emotion CNN Algorithm' button to train CNN with audio features and to get below output

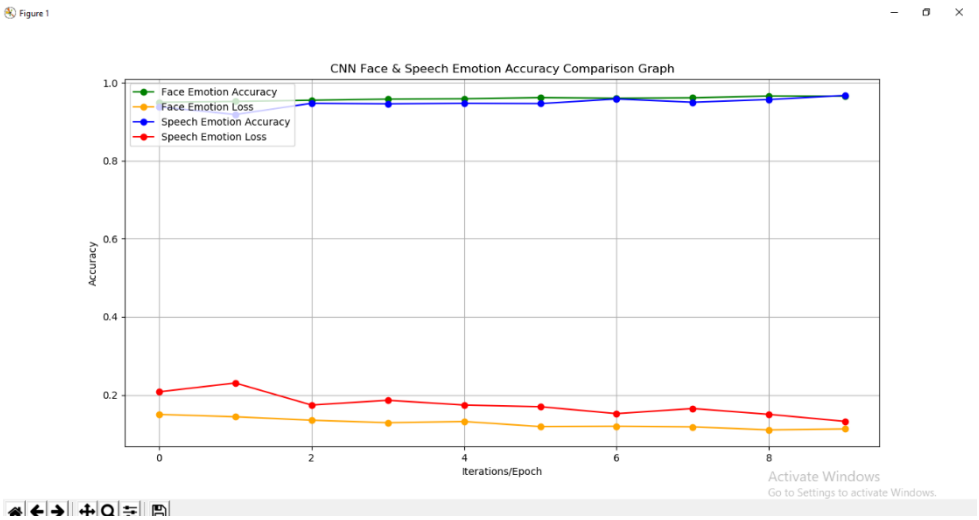

Fig. 6: Shows the Performance metrics of CNN model.

In above graph x-axis represents EPOCH and y-axis represents accuracy and loss values and we can see both algorithms accuracy reached to 1 and both algorithms loss values reached to 0. In above graph green line represents face emotion accuracy and blue line represents speech accuracy. Now click on "Predict Facial Emotion" button to upload face image and will get below result

Fig. 7: Presents the model prediction of test image as Fearful.

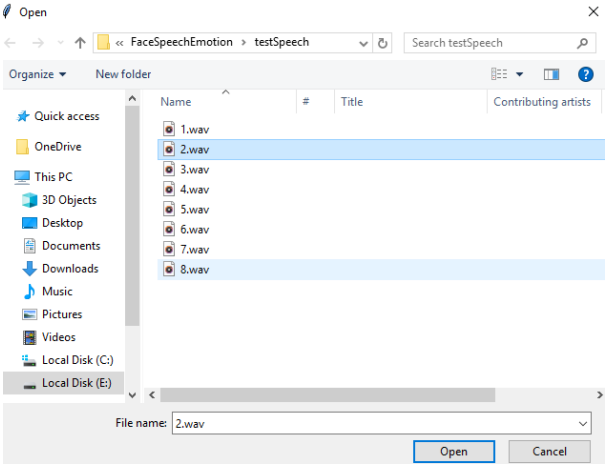

Fig. 8: Uploading the test audio for model prediction of Speech Emotion.

Fig. 9: Present the audio file emotion predicted as calm.

5. CONCLUSION

In conclusion, the development and evaluation of our multimodal emotion detection system underscore the significance of integrating advanced technologies to address the complexities inherent in recognizing and interpreting human emotions. By combining state-of-the-art techniques in speech

recognition, facial expression analysis, and video processing within a unified framework, our system demonstrates notable advancements in emotion detection accuracy and robustness. The extensive experimentation and evaluation conducted on benchmark datasets provide compelling evidence of the efficacy and reliability of our proposed CNN-based approach. Through this research, we have contributed to the advancement of affective computing by offering a scalable, adaptable, and high-performance solution for multimodal emotion detection. This represents a significant step forward in the field, with implications across various domains including human-computer interaction, virtual reality, mental health monitoring, and beyond.

REFERENCES

[1] Bjorn S, Stefan S, Anton B, Alessandro V, Klaus S, Fabien R, Mohamed C, Felix W, Florian E, Erik M, Marcello M, Hugues S, Anna P, Fabio V, Samuel K (2013) Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism

[2] Deepak G, Joonwhoan L (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-Ortiz S (2020) A machine 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 databases. Pattern Recogn 44:572–587.

[5] Eyben F, Scherer KR, Schuller BW et al (2016) The Geneva minimalistic acoustic parameter set (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 region specific features and support vector machines. Multimed Tools Appl 76:7803–7821.

[7] Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. Accessed 1 Mar 2020

[8] Hamm J, Kohler CG, Gur RC, Verma R (2011) Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. J Neurosci Methods 200:237–256

[9] Happy SL, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In Proc 4th Int Conf Intell Human Comput Interact 27–29:1–5

[10] Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. IEEE Conf Comput Vision Pattern Recognit Workshops (CVPRW).

[11] He J, Li D, Bo S, Yu L (2019) Facial action unit detection with multilayer fused multi-task and multi-label deep learning network. KSII Trans Internet Inf Syst 7:5546–5559.

[12] Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio–visual emotional big data. Inf Fusion 49:69–78.

[13] Hutto CJ, Eric G (2014) VADER: A parsimonious rule-based model for sentiment analysis of social media text. AAAI Publications, Eighth Int AAAI Conf Weblogs Soc Media

[14] Iliou T, Anagnostopoulos C-N (2009) Statistical evaluation of speech features for emotion recognition. In: Digital telecommunications ICDT'09 4th Int Conf IEEE 121–126

[15] Jia X, Li W, Wang Y, Hong S, Su X (2020) An action unit co-occurrence constraint 3DCNN based action unit recognition approach. KSII Trans Internet Inf Syst 14:924–942.

[16] Joseph R, Santosh D, Ross G, Ali F (2015) You Only Look Once: Unified, Real-Time Object Detection arXiv preprint arXiv:1506.02640

[17] Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. 2015 IEEE Int Conf Comput Vision (ICCV).

[18] Kao YH, Lee LS (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: InterSpeech

[19] Kaulard K, Cunningham DW, Bülthoff HH, Wallraven C (2012) The MPI facial expression database—A validated database of emotional and conversational facial expressions. PLoS One 7:e32321.

[20] Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recogn Lett 34:1159–1168.

[21] Ko BC (2018) A brief review of facial emotion recognition based on visual information. Sensors 18.

[22] LeCun Y, Bengio Y, Hinton G (2015) Deep learning, Nature 521.

[23] Lee C, Lui S, So C (2014) Visualization of time-varying joint development of pitch and dynamics for speech emotion recognition. J Acoust Soc Am 135:2422.

[24] Li S, Deng W (2020) Deep facial expression recognition: A survey. IEEE Trans Affective Comp (Early Access).

[25] Liu M, Li S, Shan S, Wang R, and Chen X (2014) Deeply learning deformable facial action parts model for dynamic expression analysis. 2014 Asian Conference on Computer Vision (ACCV) 143–157.

[26] Lotfian R, Busso C (2019) Curriculum learning for speech emotion recognition from crowdsourced labels. IEEE/ACM Trans Audio, Speech Lang Processing 4.

[27] Luengo I, Navas E, Hernáez I, Sánchez J (2005) Automatic emotion recognition using prosodic parameters. In: Interspeech, 493–496.

[28] Ma Y, Hao Y, Chen M, Chen J, Lu P, Košir A (2019) Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. Inf Fusion 46:184–192.