History of Medicine, 2020, 6(1): 564-572

DOI: 10.48047.v6.1.2020.5

Multimodal Emotion Recognition using Convolutional Neural Networks for Advanced Affective Computing and Human-Computer Interaction

Dr. S. Sreenath Kashyap[1], Dr. U. Rajender[2]

[1]Associate Professor, Department of Electronics and Communication Engineering, Kommuri Pratap Reddy Institute of Technology, Hyderabad, Telangana. sreenathkashyaps@gmail.com

[2]Assistant Professor, Department of Electronics and Communication Engineering, Vaageswari College of Engineering, Karimnagar, Telangana. urajender10@gmail.com

ABSTRACT

The precise identification and interpretation of human emotions are essential in the contemporary landscape of affective computing and human-computer interaction. This document outlines an advanced multimodal emotion detection system that integrates the latest techniques in facial expression analysis, speech recognition, and video processing. Traditional methods for emotion identification exhibit limitations, particularly in their ability to accurately capture the intricate and dynamic emotional states of individuals. This study addresses the identified challenges by developing a comprehensive framework that integrates multiple modalities to enhance the accuracy of emotion identification. An extensive analysis of current techniques, including feature-based and rule-based systems, reveals significant drawbacks such as limited scalability and an inability to handle complex emotional expressions. This document introduces a new method utilizing Convolutional Neural Networks (CNNs), motivated by the need for enhanced efficiency and adaptability in emotion recognition systems. Convolutional Neural Networks (CNNs) provide advantages through hierarchical representation and automatic feature learning. This facilitates the extraction of discriminative emotional cues from various modalities, including speech, facial expressions, and video data. The proposed model aims to address the limitations of traditional methods by employing convolutional neural networks (CNNs) to enhance the reliability and accuracy of emotion recognition across diverse environments and situations.Comprehensive testing and evaluation on benchmark datasets demonstrate that our multimodal CNN-based approach is proficient in accurately identifying and classifying a wide range of emotional states. This research contributes to the field of affective computing by providing a scalable, flexible, and high-performance solution for multimodal emotion recognition, with potential applications in virtual reality, human-computer interaction, and mental health monitoring, among other areas.

Keywords:Emotion recognition,Human-Computer Interaction, Multimodal Emotion Detection,Convolutional Neural Networks.

1. INTRODUCTION

Recognizing and interpreting human emotions is essential for effective communication and interaction. This capability is critical across multiple domains, including human-computer interaction, virtual reality, healthcare, and marketing. Emotion detection systems have

attracted significant attention because of their ability to improve user experiences, tailor services, and offer important insights into human behavior. Traditional methods for emotion detection frequently do not adequately capture the complexity and subtlety inherent in human emotional expressions. The development of advanced multimodal systems is required to integrate information from various sources, including speech, facial expressions, and body language, in order to enhance the robustness and nuance of emotion recognition. The study of FER has received significant attention in recent decades due to the swift advancements in artificial intelligence methodologies. Numerous feature-based methods have been investigated for FER systems. The methods identify a facial region within an image and extract geometric or appearance features from that region. The geometric features typically encompass the interrelationships among facial components. Facial landmark points serve as representative examples of geometric features [2, 30]. Appearance features are derived from the global facial region characteristics or various types of information pertaining to facial regions [20]. The global futures typically encompass principal component analysis, local binary pattern histograms, among other methodologies. Multiple studies segmented the facial area into distinct local regions and extracted appearance features specific to each region [6, 9]. The significant regions within these local areas are initially identified, leading to an enhancement in recognition accuracy. In recent decades, the extensive development of deep-learning algorithms has led to the application of convolutional neural networks (CNN) and recurrent neural networks (RNN) across various fields of computer vision. The CNN has demonstrated significant performance in multiple studies, including face recognition, object recognition, and facial emotion recognition (FER) [10, 16]. Despite the superior performance of deep-learning-based methods compared to conventional approaches, challenges persist with micro-expressions, temporal variations of expressions, and other related issues [21].

Speech signals represent a fundamental medium for human communication, characterized by their ease of measurement in real-time. Speech signals encompass both linguistic content and implicit paralinguistic information, which includes emotional cues related to the speakers. Unlike FER, the majority of speech-emotion recognition techniques focus on the extraction of acoustic features, as end-to-end learning methods, such as one-dimensional CNNs, are less effective in automatically extracting relevant features when compared to acoustic features. Consequently, the integration of suitable audio features is essential. Numerous studies have established a correlation between emotional vocalizations and their corresponding acoustic features [1, 5, 14, 18, 27]. Due to the absence of explicit and deterministic mapping between emotional states and audio features, the rate of recognition for speech-based emotion recognition is lower compared to other methods, such as facial recognition. Identifying the optimal feature set is essential in the process of speech-emotion recognition. The integration of speech signals and facial images enhances the accuracy and naturalness of emotion recognition by computers. The integration of emotional information must be executed with precision, ensuring appropriate combinations across varying degrees. Multimodal studies primarily concentrate on three strategies: feature combination, decision fusion, and model concatenation. Deep-learning technology, applicable across various fields, can play a key role in the combination of multiple inputs [7, 22]. Model concatenation provides a straightforward method for integrating models with varying inputs. Models

process various data types to produce corresponding encoded tensors as output. The tensors associated with each model may be interconnected through the use of the concatenate function. Yaxiong et al. transformed speech signals into mel-spectrogram images, enabling a 2D CNN to process the images as input. Furthermore, the facial expression image is processed using a 3D convolutional neural network (CNN). Following the concatenation of the two networks, a deep belief network was utilized for the complex nonlinear fusion of multimodal emotion features [28].

2. 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.

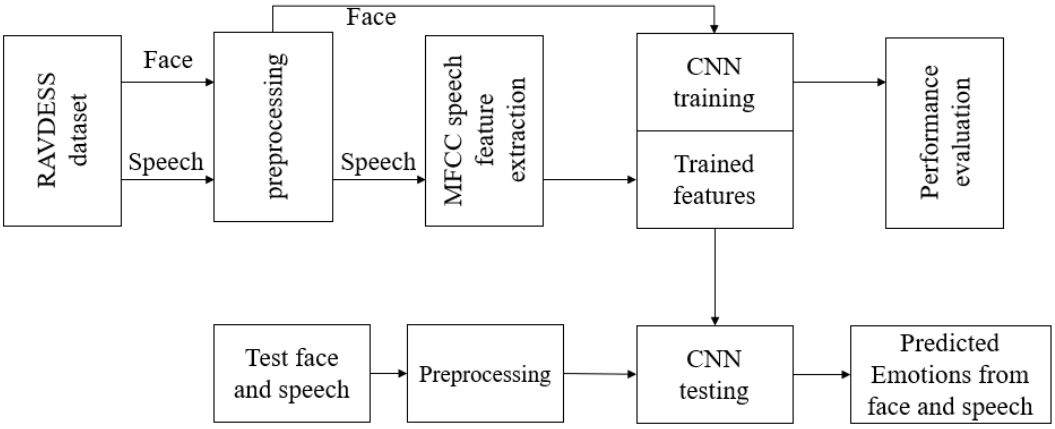

Figure 1: Proposed block diagram

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.

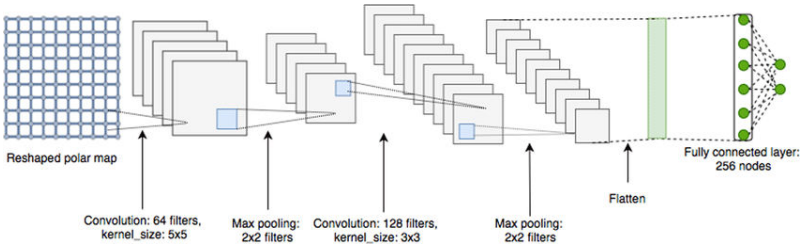

Fig. 2: Architectural of CNN model.

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.

3. 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.

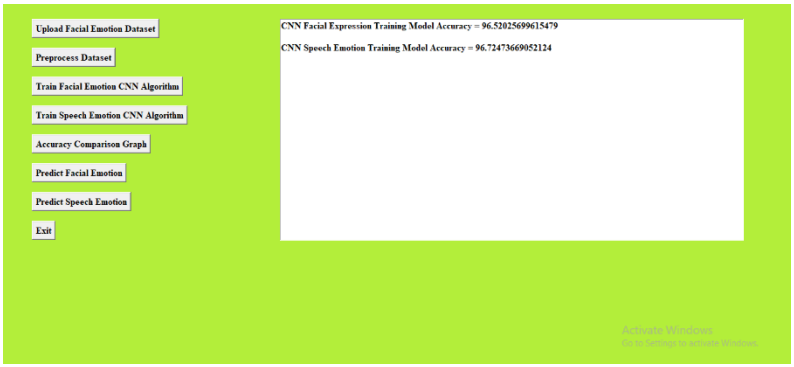

Fig. 4: Shows the Accuracy of CNN model for Speech Emotion.

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

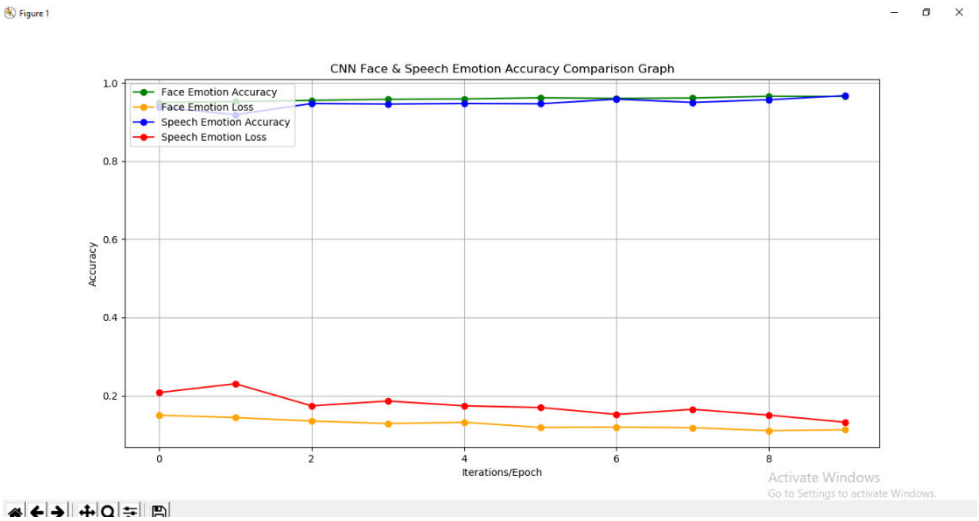

Fig. 5: Shows the Performance metrics of CNN model.

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

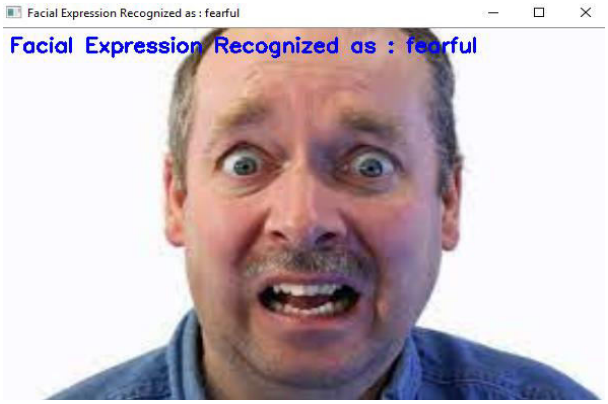

Fig. 6: Presents the model prediction of test image as Fearful.

Fig. 7: 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.