RECOGNITION OF MULTIMODAL EMOTIONS WITH A HIERARCHICAL NEURAL NETWORK

[1]Tanvi Sharawat*, [2]Dr. V.K. Srivastava

[1]Research Scholar, Department of Computer Science and Applications
Baba Mastnath University, Asthal Bohar, Rohtak, Haryana, India
[2]Supervisor, Professor and Head, Department of Computer Science and Applications
Baba Mastnath University, Asthal Bohar, Rohtak, Haryana, India
Email ID: *tsharawat3@gmail.com*

Accepted: 11.03.2022 **Published**: 01.04.2022

Keywords: Deep Learning, Electroencephalogram, Multimodal Emotion Recognition, Multiscale Features.

Abstract

Deep learning together with electro-encephalograms has been extensively used in recent years in the field of multimodal emotional recognition. Because of the complexity of electroencephalograms, some scientific researchers have employed profound education to uncover new parts of emotional detection. In previous experiments, the neural model network was used to automatically extract functionalities and fully recognise the feeling and achieve particular results. However, it is still being studied with a convolutionary neural network to extract hierarchy features. The paper therefore proposes hierarchical neural network fusions to explore data possible data information through the building of diverse structures in the hierarchy of the network, the extraction of multi-scaling features and the use of fusion to combine weights with statistical features manually extracted to produce the final vector characteristics. In this study, the DEAP and MAHNOB-HCI data sets' valence and excitation dimensions are examined in binary classifications in order to test the model's effectiveness. The results show that 84.71% and89.00% of the two data sets were precise with the proposed model, which indicates that the model provided in this study was higher than prior models for classification of deep emotional learning in function extraction and fusion.

Paper Identification

**Corresponding Author*

1. Introduction

Emotional recognition has gained more attention in numerous fields in recent years and is an important factor for the implementation of human-computer interaction systems. Emotions are complex physiological processes linked to various external and internal activities. Emotions are difficult. Emotional expressions include physiological reactions such as skin temperature and heart speed, and non-physiological reactions such as facial expressions and linguistic expressions[1]. The extraction and fusion of neural multimodal affective hierarchical modular networks are key procedures for the discovery and

execution of the weaknesses of functional fusion and decision-taking fusion are addressed and neural system performance improvements through the integration of data from various patterns between the sub-modulus of each module[2]. The model employed RFs as the classifier to calculate the cardiovascular and galvanic skin response time domain and frequency domain characteristically properties and to repeatedly remove and select features and support vector machines[3]

Multimodal Emotion Recognition Using HCNN entropy

This structure is assessed by quantitative entropy (EEG) in order to exclude constant entropy values, which change over time to reach emotionally irrelevant recognition, and the best average accuracy rate is 85.11%. These processes can provide good results but, due to manual removal and early fusion of EEG data, duplication and loss of essential functions are achieved in numerous modalities. There is therefore still much study on how functions might be extracted and calculated efficiently[4]. While scientists have developed several EEG extraction and fusion methods, these methods have issues including excessive time complexity and poor accuracy. To deal with these challenges, numerous deep learning models in the field of emotional recognition have been widely suggested[5] These approaches allow CNNs to detect powerful spatial patterns in images and RNNs to extract time-speaking and video data for classification, as well as enabling AEs to learn uncontrolled features[6]. Every CNN layer has some traits that reflect crucial information at its own level of abstraction. The first layer extracts local and regional characteristics as the process of convolution progresses layer by layer and the last layer extracts only global characteristics. Multi-modal identification of emotions attempts to combine predictive capabilities and physiological features of individual behavioural pathways for a precise classification[7].

The multimodal emotional recognition issues are:

• Combining and modelling of several modal data is more challenging than unimodal emotional recognition systems.

• A high degree of predicted accuracy is necessary, even in the context of multimodal emotion recognition, which demands procedures and processes for extraction and fusion[8].

This study presents a new hierarchy of fusion convolution neural networks (HCNN) using HCNN information for the multi-modal signal representation and combining the statistical characteristics of the time and feature to imprint the layered incriminatory architecture by establishing variable convolutionary kernel sizes and numbers on the convolutionary strata of the CNN.

A hierarchical fusion convolution neural network model based on the multimodal feature extraction is developed. The results of this research are presented as follows.

• Weights are utilised to fuse hierarchical convolutionary functions into a global vector based on the hierarchical network structure within a CNN.

• In order to strengthen the emotional recognition system, the physiological signals and emotions are employed in several video segments[9].

2. Literature Review

Emotional recognition research suggests that two types of emotional awareness exist: first the detection of physiological signals while second the observation of emotional behaviour. Because people are not capable of evocating the actual emotional state of humans, emotional detection based on the behaviour has limitations. Their current technology is based on an active consciousness. Therefore, research using physiological indicators to recognise emotions garnered more attention.

2.1 Related works

1. **Bao et al (2020)** “Several EEG data and eye movements have been used to study the

influence of gender differences and to evaluate the gender differences of five emotions using two neural network classifiers. The results demonstrated a rather high overall accuracy of the same-sex techniques.

2. **Chen et al (2019)** suggested a multi-channel EEG extraction approach which combines differential entropy with linear discrimination and attained an 82.5 percent accuracy rate.
3. **Huang et al (2019)** proposed a framework for multimodal emotional recognition combining facial expressions with EEG inputs. This framework uses a series of vector machine support classifications to detect and merge fusion technology with valence and exciting learning objectives. The results demonstrate that after fusion the effect is considerably better than that of a single form and the best outcome is 70%.
4. **Xing et al (2019)** The average DEAP accuracy rate is 81.10%; the EEG emotion recognition framework is a mixed linear EEG and emotional time model that provides a multi-channel framework by decomposing EEG-source signals from collected EEG signals and increasing classification accuracy using contextual correlations of the EEG sequences.
5. **Granados et al (2019)** The CNN has been used to automatically extract characteristics from a range of physiological inputs and to anticipate sensations via fully connected network layers. These studies show that this technique has been more precise in the emotional classification. Studying multimodal emotional recognition glasses' application possibilities,
6. **Guo et al. (2021)** Used images of Eye, Eye and EEG to grade five emotions. The results showed the three ways that the five additional emotions can become better understood. Furthermore, the results show that the classifier can achieve accuracy by using fusion characteristics of human photos and eye movements of 71.99 percent."

3. Research Methodology

Figure 3.1 shows the HCNN-based multimodal recognition model. Given that no single signal can be utilised to classify emotion categories properly, the tracking of actual human emotions is carried out using multiple observes of EEG data and peripheral physiological signals (PPS). The HCNN is used to extract multimodal signals and weighted fusion is utilised to enhance the multiscale expression of multimodal signals that improves emotional recognition. This paper will describe in full below the emotion recognition model proposed.

3.1 Data pre-processing and observation

In the document people who have been invited to observe selected video clips of different emotions and collect their symptoms by sensors at the same time, are utilised to produce human emotions. "Video-stimulation The middle thirty's of each film, including a galvanic skin reaction (GSR) a bread band (RESP), a skin temperature (TEMP) and a plethysmograph, is taken as experimental data by six satellites (FP1, FP2, AF3, AF4, F3, F4) and four PPS signals for inputs into the emotional recording model. The EEG and PPS signal is supplied with a band-pass filter and low-pass filter to prevent noise and to test signals. The median 5 is employed during the experiment in two separate categories: high value (HV)/low value (LV) and high excitation (HA)/low excitement, according to the valences and excitement values of the films. The experiment is divided into two different classes (LA). The problem of the unmatched number of samples in the experiment categories increases sample points to increase the number of samples in the smaller categories to equate the data set (see section IV for the

specific procedure of expansion). Enhance standardisation capacity of the categorization model. The final sample size appears in Table 3.1. The EEG and PPS vectors will be combined into one single vector as an HCNN input and functional calculation after sample data have been processed into a set of data using the observation-level function fusion procedure. It then extracts the hierarchical convoluted function."

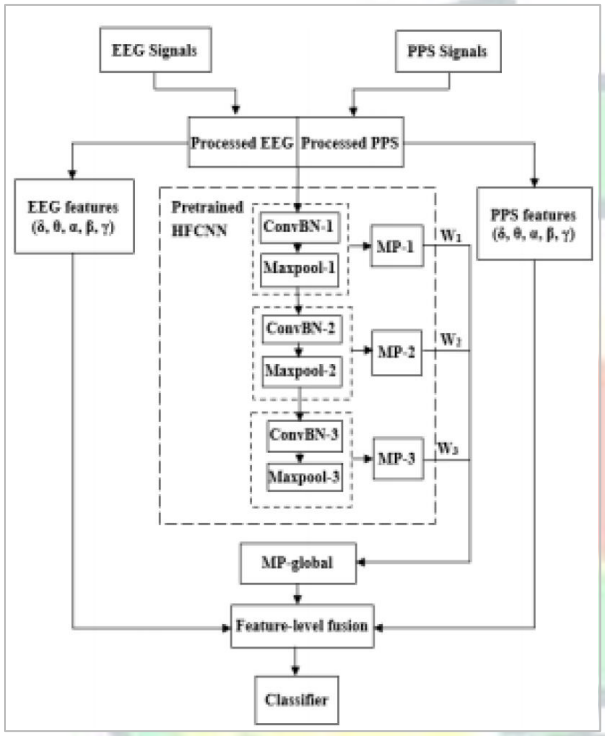

Figure 3.1 "Multimodal emotion recognition model using the hierarchical fusion convolutional neural network.(HCNN)"[1]

multimodal representation of multimodal information. The performance of the model is significantly improved. Figure 3.2 shows the particular process. In a uniform vector at the observatory level, the multimodal signals are first merged to achieve the HCNN layer input. In the HCNN model structure the characteristics of emotional data are also based on three progressive procedures to draw hierarchically convolutionary features. The second is made up of 12 5 size filters and the third is made up of 16 3 to 3 size filters. The second consists of twelve size 5 filters. In the suggested concept, the pooling layers use maximum pooling. The maximum pooling is feasible in 2 / 2 phases. The multimodal signal observations enter three layers in step and translate the maximum layer into MPs-1, MPs-2 and MPs-3 output functionality that can be represented in the Figure. 2. The results of the overlay feature of the previous layer can be accessed with the Relu Activation function. Those features are removed after maximum group layers so that they can be compressed without significant information being sacrificed. With each of three procedures with convolutionary layers, the convolutionary features have differed. Weights are ultimately merged to form weights from the three processes of increase for the MPs-1, MPs-2 or MPs-3 functions. The CNN model is applicable before function extraction.

DEAP dataset				MAHNOB-HCI dataset			
Label	Data Quantity	Label	Data Quantity	Label	Data Quantity	Label	Data Quantity
HA	33930	HV	32580	HA	11475	HV	12015
LA	33670	LV	32020	LA	11025	LV	11985
Total	67600	Total	64600	Total	22500	Total	24000

Table 3.1 "the number of expanded sample points"

3.2 Hierarchical convolutional neural network (HCNN)

The multimodal input indicators provide hierarchical characteristics to address the lack of essential features on the conventional neural networks and to create a

The pattern and weights are then frozen and the melting pattern is trained after training. Maximum pools in this article are used to optimise convolutionary features, minimise the next layer settings and avoid overlaps.

We use the method for current experimental progress in CNNs in deep learning and machine learning. A drop-out layer is implemented before the incremental process convolutions are overrun. There is also a supervised learning strategy, which translates every input model into an output class. The target group is the softmax layer, which generates prediction values. The reverse propagation technique optimises the network parameter based on the error between the anticipated value and the actual value, a tiny stochastic drop in batch sizes is performed.

4.1. Data Gatherings

This study tests the model in two sets of data, which contain several physiological signals and emotional evaluations. 32 subjects saw 40 films of varying moods during the DEAP data collection experiment, each of which occupied approximately a minute. Several physiological signals and several 1-9 metrics have been recognised.[11]. Because brain areas associated to the front-local area are highly accurately recognised, data collected amounts to 128 Hz. Includes: alpha (13-30 Hz), beta (30-33 Hz), delta (8-13 Hz) and gamma

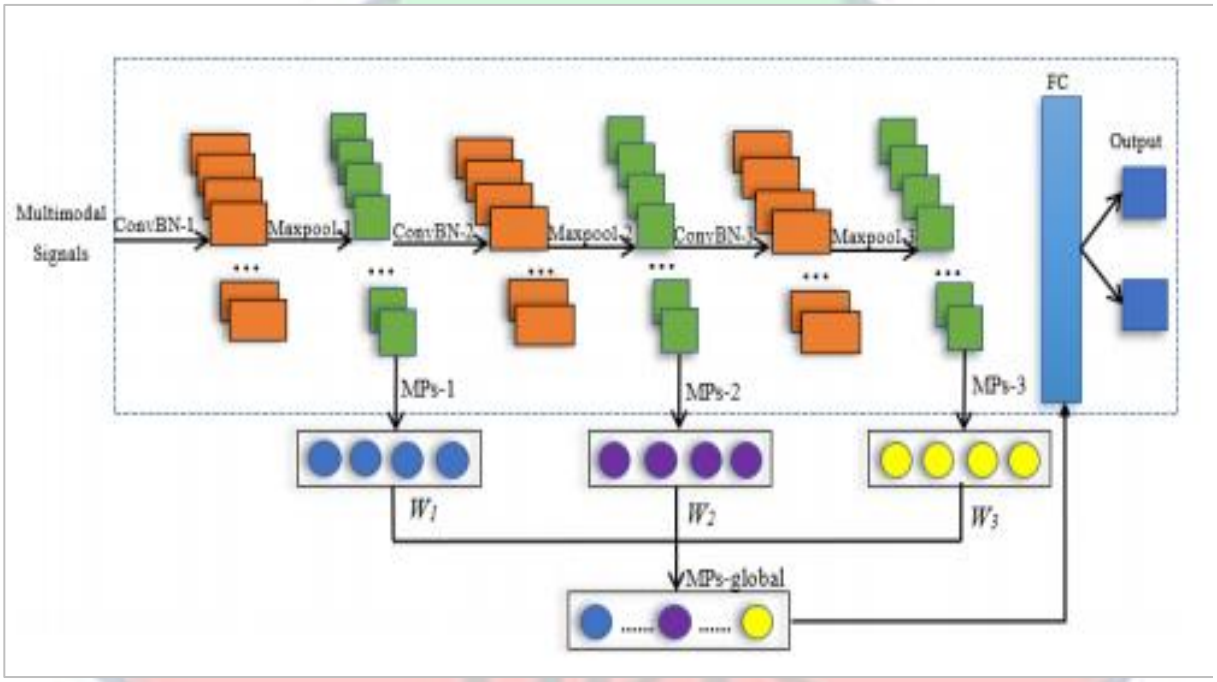

Figure 3.2 "The process of the HCNN as per this article"[10]

4. Results and Analysis

This study focuses on multimodal physiological signals extraction and fusion approach. We have experimented with DEAP and MAHNOB-HCI data sets to evaluate the efficiency of our suggested technique. The local hierarchical convolution features recovered from the HCNN finally provide a global function vector with weights in combination with manual features. RF is utilised using 10-times cross-validation procedures to train and test the two dimensions of excitation and valence.

(4-8 Hz). Five of them are removed (4-43 Hz). Due to the mistake of the first 3 s of the video, the film is taken as experimental data in the middle of the other 30 s. Each one has 128 points of time per 6 channels, with 30 sizes of data (video times). A 6-s window with 50% overlay is used to resolve the problem of insufficient sample points in depth to raise the number of test points and apply them to each band of the samples[12].

For example, when the overlapping window was raised to 3 s and there were 40 = 360 signal segments per topic in 40 tests for every topic, then the samples point

number in each of the 40 experiments was 360 = 1800 for each subject after a window for each topic. The most recent measures are 4608 = 6 = 6 (visual times) (epochs) [13]. for every participant. The threshold is separated into two categories in the aspects of valence and excitement. “The mean value is 5. Where a video has an arousal or valence value more than or equivalent to 5, it is HV/HA, otherwise it is LV/LA. The size of the label for each subject is therefore 1 to 1800. 30 individuals (17 females and 13 males) have produced the MAHNOB HCI data set to view 20 videos with various emotions and EEG signals, peripheral physiological signals and movements of the eyes. The database draws on the twenty videos of famous films, which emotionally focus on themes. Every clip lasted between 34 and 117 seconds”[14].

“Dimensions	Arousal		Valence	
Dataset	DEAP	HCI	DEAP	HCI
CNN	75.98%(RF450)	7.17%(RF600)	75.98%(RF200)	71.28%(RF500)
HCNN	83.28%(RF450)	88.28%(RF600)	84.71%(RF200)	89.00%(RF200)”

Table 4.1 “the experimental results under the optimal parameter settings”[15]

5. Multimodal Emotion Recognition Results

In addition, fusion characteristics are included in both data sets in the RF classification model in order to remove the effects of classification parameters on experimental results, to analyse the effect on the accuracy of radiofrequency alterations. Figure 4.1 illustrates the curve changes in accuracy of DEAP and MAHNOB-HCI RF dimensions with an ever-increasing number of tree dimensions. The figure shows that following processing, the exactness of multimodal functions rises gradually, by making more RF decisions, and finally achieves a certain degree of stable precision. For the two sets of data of various sizes and parameters, we have therefore determined specific values. There are 450 and 200 DEAP Decision Treaties and there are 600 and 500 MAHNOB-HCI decisions. Table 4.1 shows the matching medium accuracy rates. The DEAP data set shows an average exciting precision of 83.28% and an estimated 84.71% in Table 2. In the MAHNOB-HCI data set the arousal accuracy is 88.28% and in the rating is 89.00%. You may deduce that the value performance in the parameter optimization scenario is higher than the excitement with the same data set.

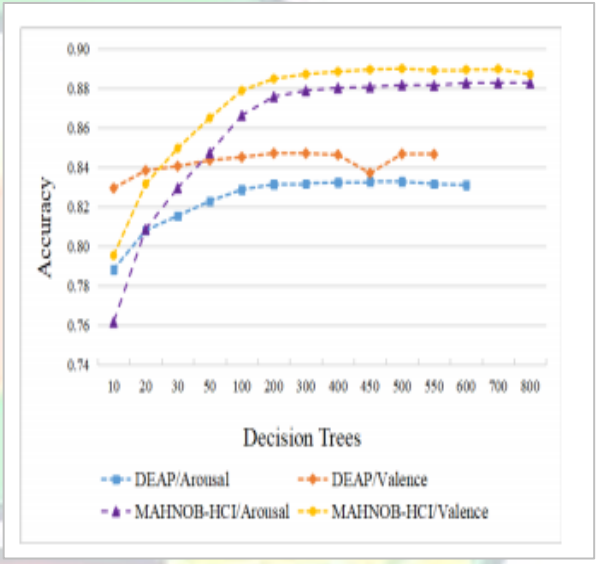

Figure 4.1 **“**The change curve of accuracy with the parameter setting in the classifier”[12]

5. Conclusion

This research is a new multi-modal hierarchy for early fusion learning within the context of emotional recognition. As indicated earlier, MPs-1, MPs-2 and MPs-3 have received the global fusion. For each participant in the HCNN model to perform binary valence and exciting categorization trials, a tenfold cross-validated dataset has finally been constructed. The experimental results show that the combination of signals and HCNN extracting characteristics provides greater and steady performance, overcoming the absence of significant characteristics and increasing the multimodal representation of traditional CNN. By

analysing HCNN features and statistical properties, we have discovered that substantial accuracy is achieved using several strategies. In order to prevent the technical process of manual removal and selection before a traditional master classification, we introduced a subject-independent emotional recognition system that enables us to realise the real emotional situation and effectively improve the accuracy and stability of multimodal identity.

RÉFÉRENCIAS

[1] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Futur. Internet*, vol. 11, no. 5, p. 105, 2019.

[2] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[3] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep Learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Futur. Gener. Comput. Syst.*, vol. 101, pp. 542–554, 2019.

[4] H. Zhang, "Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder," *IEEE Access*, vol. 8, pp. 164130–164143, 2020.

[5] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 176–183.

[6] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," *IEEE Access*, vol. 7, pp. 18940–18950, 2019.

[7] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2909–2917.

[8] M. Riyad, M. Khalil, and A. Adib, "Cross-subject EEG signal classification with deep neural networks applied to motor imagery," in *International Conference on Mobile, Secure, and Programmable Networking*, 2019, pp. 124–139.

[9] D.-W. Chen *et al.*, "A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition," *Sensors*, vol. 19, no. 7, p. 1631, 2019.

[10] H. Huang, Z. Hu, W. Wang, and M. Wu, "Multimodal emotion recognition based on ensemble convolutional neural network," *IEEE Access*, vol. 8, pp. 3265–3271, 2019.

[11] S. Koelstra *et al.*, "Deap: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2011.

[12] D. Y. Choi, D.-H. Kim, and B. C. Song, "Multimodal attention network for continuous-time emotion recognition using video and EEG signals," *IEEE Access*, vol. 8, pp. 203814–203826, 2020.

[13] J. Zhu, X. Zhao, H. Hu, and Y. Gao, "Emotion recognition from physiological signals using multi-hypergraph neural networks," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 610–615.

[14] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2011.

[15] D. Wu, J. Zhang, and Q. Zhao, "Multimodal

fused emotion recognition about expression-EEG interaction and collaboration using deep learning," *IEEE Access*, vol. 8, pp. 133180–133189, 2020.