Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Comparative analysis on Emotion Recognition by Multi-Channel CapsNet Learning Framework

D. Vinod Kumar

Professor & Head, Department of Biomedical Engineering, Vinayaka Mission's Kirupanada Variyar Engineering College, Vinayaka Mission's Research Foundation (Deemed to be University), Salem, India

E-mail: vino.kd@gmail.com

Abstract

This study uses electroencephalography (EEG) data to construct an emotion identification system utilizing a deep learning model. Modeling numerous data inputs from many sources, such as physiological signals, environmental data and video clips has become more important in the field of emotion detection. A variety of classic machine learning methods have been used to capture the richness of multimodal data at the sensor and feature levels for the categorization of human emotion. The proposed framework is constructed by combining the multi-channel EEG signals' frequency domain, spatial properties, and frequency band parameters. The CapsNet model is then used to identify emotional states based on the input given in the first stage of the proposed work. It has been shown that the suggested technique outperforms the most commonly used models in the DEAP dataset for the analysis of emotion through output of EEG signal, functional and visual inputs. The model's efficiency is determined by looking at its performance indicators.

Keywords: CapsNet, emotion analysis, EEG signal, classification, denoising approach, speech processing.

1. Introduction

A rise in the popularity of sensors and low-power integrated circuits has led to the creation of wearable devices that can gather and transmit real-time data for a longer duration. Human physiological sensing and other natural environmental elements, such as weather and noise levels, may be combined with these data sources to create new and inventive methods for detecting human activity [1]. Whether aware or unaware, humans have an emotional response to an event or situation. Various biological and physical responses, such as speech,

DOI: https://doi.org/10.36548/jaicn.2022.2.003
Received: 19.04.2022, received in revised form: 16.05.2022, accepted: 30.05.2022, published: 14.06.2022

language, gestures, facial expressions, and bio signals, are used to represent the mood. Emotions can't be hidden, even if people don't want to show them. As a result, the ability to respond to one another's emotions is vital for effective communication [2-4].

Many applications of human-computer interaction rely heavily on emotional responses. Systems that respond to user requests without explicit requests or instructions have recently been introduced by several companies. By detecting the emotions of its users, the systems are able to pick up on the implicit demands of their users. Recently announced television sets automatically adjust their contrast and brightness based on what's being shown on screen. For example, since the user's emotions are heightened while the television plays football as its content, its contrast and brightness rise.

There are a variety of ideas and concepts for identifying and distinguishing various states of emotional in cognitive manner. There was an ongoing discussion over how to define emotions, and one claimed that emotions may be categorized by recent deep learning model that established the embedded intelligence connections [5, 6]. There are four emotional states found in this bi-dimensional paradigm, which comprises of,

1. High stimulation high valence,
2. High stimulation low valence,
3. Low stimulation high valence
4. Low stimulation low valence (not in state)

As a result, the valence-arousal paradigm may be used to represent and analyse almost any typical emotional state. The extracted features are the principal components for the technique used to detect emotional content in spoken words [9]. Researchers have developed a variety of features in speech processing, such as,

1. Real source excitation features from words,
2. Special prosodic components,
3. Unwritten adhesion elements.

Linear and nonlinear classifiers are used to categorise features in the second step. Bayesian Networks (BN) or the Maximum Likelihood Principle (MLP) and Support Vector Machines (SVM) are the most often used linear classifiers for emotion identification. Usually, the voice signal is termed non-stationary. Because of this, nonlinear classifiers are believed to be successful in SER. The Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM) are two of the numerous nonlinear classifiers available for SER. The

categorization of information produced from fundamental traits is often carried out using these [10, 11].

2. Literature Survey

Many wearable gadgets like smart watches and wristbands have recently been outfitted with sensors that can constantly monitor human physiological signals (e.g., heart rate, movement and location data) as well as ambient environmental information (e.g., temperature and humidity levels, noise, brightness, etc.). A broad range of research disciplines, including healthcare and smart cities, are impacted by this. New tools and methodologies are needed for healthcare researchers to cope up with the huge multidimensional datasets that are generated by on-body and ambient sensors. These data, especially in ubiquitous and mobile computing, have been the subject of several studies over the last few decades by academics from a variety of professions [12].

Three-dimensional video data may be represented using the 3D SIFT descriptor, which was suggested by Scovanner et al. [13]. Video data may be better described using their method, which identified links between Spatio-temporal terms. It is possible to extract emotions from face data included in video sequences by using a new local descriptor that uses histograms of oriented 3D spatiotemporal gradients [14]. With their descriptor, they demonstrated that it outperformed advanced approaches in the selection of action datasets.

According to Coan et al. [15], the left and suitable frontal brain areas are linked to positive and negative emotions. Intelligence movement drops supplementary in the forward section of the intelligence than in other parts of brain activity. Signal processing, feature extraction, and classification are all necessary steps in digital voice emotion recognition systems [16]. The denoising and segmentation are examples of acoustic pre-processing [17]. Next, feature extraction is used to find the movement's most significant characteristics. To complete the process, classification algorithms must identify the appropriate emotions from the feature vectors.

2.1 Motivation of this Research

It is difficult to tell which characteristics are significant to a specific job and which are just noise. As a consequence, the ability to choose parts from a large feature set is crucial, and processing these features will need further dimensional reduction approaches. The process of extracting and selecting features is computationally time-consuming as well. It is

possible that the computational cost of feature selection will rise as the number of features grows. Many search methods may fail to find the best feature sets for a particular model. Abstract representations of sensor data that are invariant to local data changes are crucial for the complexity of human emotion recognition. It is a big issue in pattern recognition, which involves learning components that remain constant throughout time, to acquire such consistent characteristics.

3. Methodology

The raw EEG data and self-reports of emotions from the DEAP dataset are utilized in this study. The sampling rate of the EEG data is reduced to 128 Hz. The electrooculogram (EOG) effects are eliminated, and cut-off frequencies are used in band-pass filtering. The DEAP is used to test the suggested emotion recognition framework's performance [18]. The DEAP dataset includes data from 32 participants who sat through 40 music videos and had their body measurements and movies videotaped. The simple arrangement of the proposed framework has been shown in figure 1.

3.1 Filter Design

This design section comprises based on signal strength and states of the activity. The average mean reference approach is used to decrease electrical signals present in the input domain with the help of reference values. Then the amplifier section compares the power line and external interference noises with the reference values. The reference mean for each channel is computed and each channel sample is removed. These obtained values are normalized between binary values with 0 and 1. It is used to reduce the differences between the effects.

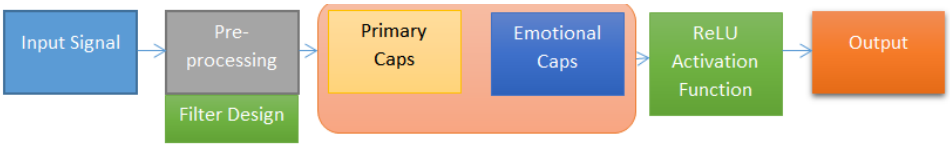

Figure 1. Multi-channel EEG analysis

3.2 Capsules Network

Classifying two-dimensional objects using a neural network is the most often used technique. However, the data routing mechanism of the CNN discards information like the object's location and posture. A network topology known as the capsule network has been developed [19] to make up for CNN's inadequacies. There are nodes and capsules in the

CapsNet model. In a capsule, neurons form a cluster. The features of an object's components are represented by activation neurons. As a whole, all the capsules make up the object's general structure by choosing a single element for it. An object's components and spatial information are preserved, unlike certain deep neural networks. CapsNet, like CNN, comprises of various layers in the proposed framework. Figure 2 depicts the construction of a complex framework and the flow of data from the input layer network to output layer network.

3.3 Proposed Neural Network by CapsNet Framework

The information is lost as a result of how the CNN's max-pooling and average-pooling functions. CapsNet, on the other hand, has the ability to store this data. CapsNet has an inherent awareness of three-dimensional space and stores all the object's states and relative positioning connections. As a result, this becomes sensitive with variance in signal feature information. The values of the pixels in the appropriate area for example, indicate the components of a picture. In the same way, various parts of the brain react to different emotional states. These are the characteristics of EEG channels in various locations that correspond to the multi-channel features. Because of this, when used for multi-channel EEG-based emotion detection, CapsNet can differentiate the prominent global changes in information represented by the human brain in distinct emotional states. As an added bonus, CapsNet learns quicker and consumes fewer samples than CNN. Therefore, a feature band and CapsNet are combined to create a deep learning framework for EEG-based multi-channel emotion identification.

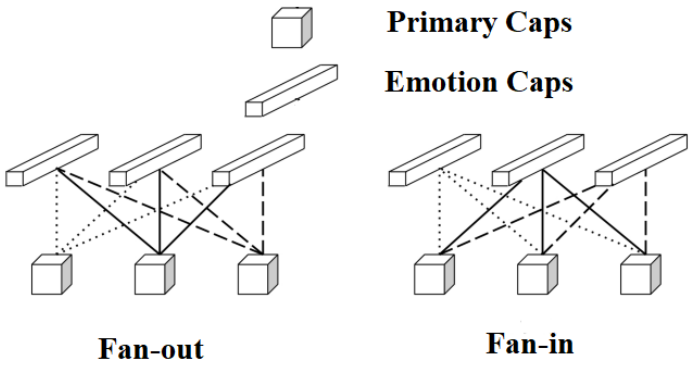

Figure 2. Path Capsule Networks

3.4 CapsNet Architecture

There are four elements in the CapsNet-based concept. First, a Rectified Linear Unit (ReLU) is used to conduct convolution operations on the input band matrix to identify local

features. The second component is the principal capsules (PrimaryCaps), which use a convolutionary method to send data to the capsules. Lastly, the dynamic routing method between capsules utilized for emotion identification is included in the EmotionCaps (emotion capsules) as shown in figure 2.

3.4.1 Hyperparameter Measures

The last step is to reassemble the input matrix using the final capsules' output. Different hyperparameter and training parameter combinations are examined to find the optimal combination for the complex neural network. In order to keep the computational complexity of the suggested model to a minimum, only the first convolution layer's parameters, PrimaryCaps filters, and EmotionCaps parameters are included in the hyperparameters.

3.4.2 Performance Measures

The performance of the capsule network is influenced by the number and type of trainable parameters. As a result, filter design that goes from positive to negative in pursuit of the best model parameter combination is used. In order to figure out how many model parameters are required, a difficult upfront decision has to be made.

4. Results and Discussion

Emotion recognition study using physiological signals and facial expressions acquired by the DEAP uses a standard database that includes physiological signals from 32 people. A great deal of work has been done on emotion recognition using the information created here for further research work. The measures for DEAP dataset are noted in figure 3.

The raw EEG data from the DEAP dataset, together with the participants' self-reports of their emotions, are utilised in this study. 512 Hz EEG waves are reduced to 128 Hz. Except for the reference findings, the electrooculogram (EOG) effects showed that the suggested technique has the greatest identification accuracy with various dimension shown in figure 4. Using the suggested strategy, it is possible to improve the performance of EEG-based multi-channel emotion identification. The various feature band and CapsNet-based framework performs a good job in identifying EEG emotions regardless the subject's emotional state. The various arousal, valence and dominance values are noted in table 1. The results are categorized with the help of arousal, valence and dominance.

Figure 3. Measures from DEAP dataset for arousal vs valence

Table 1. Computed value by various algorithms

Model	Accuracy (Average value)			Goodness to fit Description
	Arousal	Valence	Dominance	
CNN	66.25%	64.08%	63.32%	Good and variable
kNN	61.48%	63.29%	65.39%	Average
Random Forest	52.32%	50.48%	51.34%	Good
Proposed work	70.03%	69.02%	71.21%	Highest variability

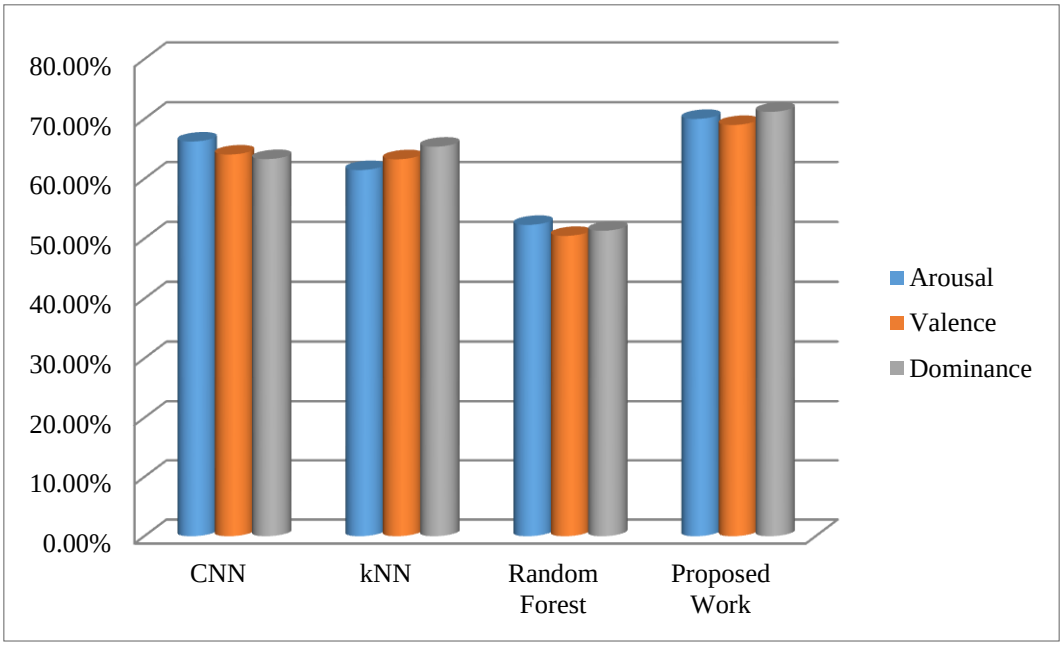

Figure 4. Accuracy performance measures from calculation

Referencing arousal and valence with dominance aspects [20-22], this study's recognition accuracy has been better. Reference [23] describes a technique for classifying samples according to the topic, although it was only used to determine which samples belonged to which category. This suggested technique has greater recognition accuracy in terms of dominance than references [24] and [21], but it is lower in the other two sizes. Without taking into account any individual variations, a framework has been devised that might better generalize the model's recognition of emotions. All individuals' EEG-based emotion detection tests performs well with the suggested framework. Figure 4 shows the accuracy performances in various domains. The above graph shows the findings of the validation set. When compared to 2D-CNN with the same network structure, CapsNet's validation accuracy improves quicker with various dynamic dimensions to attain greater accuracy.

5. Conclusion

In order to detect human emotions, this research suggests a system that makes use of multi-channel EEG framework based on CapsNet model, and hence there are two major novel components in this work. Methods for extracting features from data are first discussed in this paper. According to electrode placements and frequency band information, frequency-domain properties collected from EEG data are mapped to a multiband feature extraction. It is discovered via the use of experimental data that the suggested strategy performs admirably in terms of two key performance measures. In the future, researchers might simplify and consolidate this strategy in order to create a model that can be used to distinguish good and bad emotional states. Using this information, the specific traits that differentiate positive and negative emotional states can be identified. Wearable gadgets that can monitor a person's emotional state when they are in a resting or nearly resting position would benefit from this.

References

[1] Chen, J.; Hu, B.; Xu, L.; Moore, P.; Su, Y. Feature-level fusion of multimodal physiological signals for emotion recognition. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, Washington, DC, USA, 9–12 November 2015; pp. 395–399.

[2] Atkinson, J.; Campos, D. Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. Expert Syst. Appl. 2016, 47, 35–41.

[3] Arnau-Gonzalez, P.; Arevalillo-Herrez, M.; Ramzan, N. Fusing highly dimensional energy and connectivity features to identify affective states from EEG signals. Neurocomputing 2017, 244, 81–89.

[4] Li, X.; Song, D.; Zhang, P.; Yu, G.; Hou, Y.; Hu, B. Emotion recognition from multi-channel EEG data through convolutional recurrent neural network. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, Kansas City, MI, USA, 13–16 November 2017; pp. 352–359.

[5] Jenke, R.; Peer, A.; Buss, M. Feature Extraction and Selection for Emotion Recognition from EEG. IEEE Trans. Affect. Comput. 2017, 5, 327–339.

[6] Yin, Z.; Wang, Y.; Liu, L.; Zhang, W.; Zhang, J. Cross-Subject EEG Feature Selection for Emotion Recognition Using Transfer Recursive Feature Elimination. Front Neurorobot. 2017, 11, 19.

[7] Kwon, Y.H.; Shin, S.B.; Kim, S.D. Electroencephalography Based Fusion Two-Dimensional (2D)-Convolution Neural Networks (CNN) Model for Emotion Recognition System. Sensors 2018, 18, 1383.

[8] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio, ``EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multi-modal User Interfaces*, vol. 10, no. 2, pp. 99_111, 2016.

[9] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, ``Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 467_474.

[10] Jirayucharoensak, S.; Pan-Ngum, S.; Israsena, P. EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. Sci. World J. 2014, 2014, 627892.

[11] Khosrowabadi, R.; Chai, Q.; Kai, K.A.;Wahab, A. ERNN: A biologically inspired feedforward neural network to discriminate emotion from EEG signal. IEEE Trans. Neural Netw. Learn. Syst. 2014, 25, 609–620.

[12] Alhagry, S.; Fahmy, A.A.; El-Khoribi, R.A. Emotion recognition based on EEG using LSTM recurrent neural network. Int. J. Adv. Comput. Sci. Appl. 2017, 8, 355–358.

[13] Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional SIFT descriptor and its application to action recognition. In Proceedings of the ACM International Conference on Multimedia, Augsburg, Germany, 24–29 September 2007; pp. 357–360.

[14] Klaser, A.; Marszaek, M.; Schmid, C. A spatio-temporal descriptor based on 3D-gradients. In Proceedings of the British Machine Vision Conference, Leeds, UK, 1–4 September 2008; pp. 1–10.

[15] J. A. Coan, J. J. Allen, and E. Harmon-Jones, "Voluntary facial expression and hemispheric asymmetry over the frontal cortex," *Psychophysiology*, vol. 38, no. 6, pp. 912-925, 2001.

[16] Liu, M.; Shan, S.; Wang, R.; Chen, X. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1749–1756.

[17] Soleymani, M.; Asghari-Esfeden, S.; Fu, Y.; Pantic, M. Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Trans. Affect. Comput. 2016, 7, 17–28.

[18] Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Li, Y. Spatial-temporal recurrent neural network for emotion recognition. IEEE Trans. Cybern. 2019, 49, 839–847.

[19] Ciresan, D.; Meier, U.; Schmidhuber, J. Multi-column deep neural networks for image classification. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Providence, RI, USA, 18–20 June 2012; pp. 3642–3649.

[20] Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis; Using Physiological Signals. IEEE Trans. Affect. Comput. 2012, 3, 18–31.

[21] Salama, E.S.; El-Khoribi, R.A.; Shoman, M.E.; Shalaby, M.A.E. EEG-based emotion recognition using 3D convolutional neural networks. Int. J. Adv. Comput. Sci. Appl. 2018, 9, 329–337.

[22] Yang, Y.; Wu, Q.; Qiu, M.; Wang, Y.; Chen, X. Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In Proceedings of the International Joint Conference on Neural Networks, Rio, Brasil, 8–13 July 2018; pp. 1–7.

[23] Moon, S.-E.; Jang, S.; Lee, J.-S. Convolutional neural network approach for EEG-based emotion recognition using brain connectivity and its spatial information. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 2556–2560.

[24] Li, Z.; Tian, X.; Shu, L.; Xu, X.; Hu, B. Emotion Recognition from EEG Using RASM and LSTM. Commun. Comput. Inf. Sci. 2018, 819, 310–318.

Author's biography

D Vinod Kumar received his Bachelor of Engineering in Electronics & Communication Engineering from Periyar University, Salem, Tamil Nadu, India, Master of Engineering in Applied Electronics from Anna University, Chennai, Tamil Nadu, India. He obtained his Ph.D. Degree in Interdisciplinary in Electronics & Communication Engineering and Computer Science Engineering from Vinayaka Mission's Research Foundation (Deemed to be University), Salem, Tamil Nadu, India. He also received a Master of Business Administration from Periyar University, Salem, India. He is currently working as Professor & Head of Biomedical Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Vinayaka Mission's Research Foundation (Deemed to be University), Salem, India. He is also a member of the Board of Studies, Faculty of Engineering & Technology of Vinayaka Mission's Research Foundation (Deemed to be University), Salem, India. He has a vast experience of 17 years in teaching, research, and administration. He is currently a Member of Institution of Engineers. Life member of Indian Society for Technical Education (ISTE), Member of IEEE (USA). His current research interests include Image Processing, VLSI design, Embedded Systems, and Biometric systems. He has served as a reviewer and published many papers in peer-reviewed National & International journals and conference proceedings.