

Article

Emotion Recognition from Multiband EEG Signals Using CapsNet

Hao Chao *, Liang Dong , Yongli Liu and Baoyun Lu

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China; 211709020013@home.hpu.edu.cn (L.D.); liuyongli@hpu.edu.cn (Y.L.); luby@hpu.edu.cn (B.L.)

* Correspondence: chaohao@hpu.edu.cn

Received: 26 March 2019; Accepted: 10 May 2019; Published: 13 May 2019



Abstract: Emotion recognition based on multi-channel electroencephalograph (EEG) signals is becoming increasingly attractive. However, the conventional methods ignore the spatial characteristics of EEG signals, which also contain salient information related to emotion states. In this paper, a deep learning framework based on a multiband feature matrix (MFM) and a capsule network (CapsNet) is proposed. In the framework, the frequency domain, spatial characteristics, and frequency band characteristics of the multi-channel EEG signals are combined to construct the MFM. Then, the CapsNet model is introduced to recognize emotion states according to the input MFM. Experiments conducted on the dataset for emotion analysis using EEG, physiological, and video signals (DEAP) indicate that the proposed method outperforms most of the common models. The experimental results demonstrate that the three characteristics contained in the MFM were complementary and the capsule network was more suitable for mining and utilizing the three correlation characteristics.

Keywords: EEG signal; feature extraction; multiband feature matrix; deep learning; CapsNet; emotion recognition

1. Introduction

Emotion is an affective state of human beings accompanied by cognition and consciousness, and plays an important role in human social interactions. Human emotion recognition has become an important research focus in many fields, such as cognitive science, computer science, psychology, neuroscience, and artificial intelligence.

Emotion states can be inferred by the external and internal reactions of a human being, which change with different emotional states. Current emotion recognition research is implemented mainly according to non-physiological signals such as facial expressions [1], speech [2], body movement [3], and physiological signals like skin resistance (SR), electrocardiogram (ECG), functional magnetic resonance imaging (fMRI), electroencephalograph (EEG) [4], and magnetoencephalogram (MEG) [5]. Compared with non-physiological signals, physiological signals are not susceptible to the impact of external environment and subjective will. Therefore, physiological signals—especially EEG signals—become attractive because of the repeatability and objectivity for the estimation of emotion state. In recent years, with the development of sensor technology, it has become possible to monitor, record, and analyze multi-channel neurophysiological signals synchronously. Thus, many researchers have turned their attention to multi-channel EEG-based emotion recognition [6–12].

Various EEG features have been proposed for emotion recognition in previous studies. Time-domain analysis is implemented to obtain the characteristics of time series which describe salient information related to emotional states. The statistical features (power, mean, standard deviation, etc.) of EEG series are usually employed [13–15]. Frantzidis et al. [16] used the event-related potential (ERP)

features. In addition, other studies introduced Hjorth features (Mobility, Complexity, and Activity) [17], non-stationary index (NSI) [18] and higher-order crossing features (HOCs) [19,20] for EEG emotion recognition. Power-related measures from different frequency bands of EEG signals are often used in frequency-domain techniques [21]. The most commonly used analytical technique for EEG signals is the fast Fourier transform (FFT) [22,23]. Since the FFT cannot reflect temporal information in the frequency data, the short-time Fourier transform (STFT) is used to extract time-frequency domain features [8,24,25]. Welch's method [26] was also used to calculate the power spectral density (PSD) of different frequency bands in some studies [27,28].

The choice of specific EEG channels is critical for multi-channel EEG-based emotion recognition. Wichakam et al. [28] made use of all 32 channels as well as 10 specific channels (F3, F4, Fp1, Fp2, P3, P4, T7, T8, O1, and O2) for emotion recognition. The experimental results showed that better results were achieved when using the 10 channels compared to using all 32 channels. Jie et al. [29] proposed an emotion recognition method based on sample entropy. Their experimental results showed that the channels related to emotion states were mainly located in the prefrontal region, namely, F3, CP5, FP2, FZ, and FC2.

Many machine-learning techniques have been applied to identify emotion states, such as support vector machine (SVM) [28], k-nearest neighbor (k-NN) [30], and artificial neural network (ANN) [31]. In recent years, deep learning methods have been introduced for EEG emotion recognition [32–34]. Compared with traditional machine learning methods, deep learning methods extract features automatically based on big data and are more capable of portraying the rich intrinsic information of data. Many studies have shown the superiority of deep learning methods in emotion recognition. Kwon et al. [23] proposed a method in which time-frequency feature maps of each channel as well as other physiological features were fed into a convolutional neural network (CNN) to recognize emotion states. Li et al. [7] used frequency features to construct EEG multi-dimensional feature images. Then, a hybrid deep learning model which integrated CNN and recurrent neural network (RNN) techniques was designed to deal with the multi-dimensional feature images in the emotion recognition task.

Compared with traditional machine learning methods, deep learning has demonstrated its potential in multi-channel EEG-based emotion recognition. However, two challenges remain. Firstly, most of the attention is given to determining how to obtain the salient information related to emotional states from the time-domain, frequency-domain, and time-frequency characteristics of EEG signals. Few studies have analyzed the spatial domain characteristics of multi-channel EEG, which may also contain salient information. Further, there have been a few spatial characteristics studies limited to the asymmetry between the electrode pairs [26]. Therefore, determining how to integrate and present the common characteristics of the EEG signal with the spatial characteristics is a key problem. In addition, many deep learning methods have poor sensitivity to spatial domain characteristics in processing two-dimensional (2D) objects, such as the deep belief network (DBN) and the stacked autoencoder (SAE). Although CNNs can handle two-dimensional objects, they lack the ability to describe the relative relationship between the local parts and the whole object, which can provide useful information for classification purposes.

To address the challenges mentioned above, this paper proposes an EEG emotion recognition framework that integrates a multiband feature matrix and a capsule network (CapsNet). This is the first time that CapsNet has been employed for emotion recognition. In the framework, the PSD features are extracted from the EEG signals and multiband feature matrices (MFMs) are constructed according to the electrode positions and frequency bands. Then, MFMs are fed into a CapsNet to perform classification tasks.

The remainder of this paper is organized as follows. Section 2 describes the dataset and emotion model. In Section 3, the extraction method of MFMs and the CapsNet-based framework are introduced. Section 4 analyzes the experimental results and provides discussions. Section 5 summarizes this work.

2. Materials

2.1. EEG Dataset

The DEAP dataset was employed to validate the effectiveness of the proposed emotion recognition framework [35]. The DEAP dataset recorded several physiological signals and video signals from 32 volunteers when they were watching 40 music videos. Forty trials were implemented for each volunteer. Each trial played a specific 63-s video, with the EEG and peripheral physiological signals simultaneously recorded. After watching the videos, the subjective ratings of arousal, valence, liking, and dominance on a scale of 1–9 were gathered. EEG data of 512 Hz sampling frequency were recorded with 32 electrodes placed according to the international 10/20 system. Each electrode recorded a 63-s EEG signal, and the first 3 s was the baseline signal of the trail.

In this paper, the preprocessed raw EEG data and the corresponding emotion self-assessment in the DEAP dataset were used. EEG signals (512 Hz) were down-sampled to 128 Hz. The effects of the electrooculogram (EOG) were removed and band-pass filtering was implemented with cutoff frequencies of 4.0 and 45.0 Hz.

2.2. Emotion Model

This work relies on the valence–arousal–dominance space model because of its simplicity and ability to measure emotions relatively well [36]. Figure 1 shows the three-dimensional emotion model with valence, arousal, and dominance. The model explains emotions' change in a 3D plane, and the emotion states are determined by the values of arousal, valence, and dominance. As shown in Figure 1, the values of an emotional instance of Fun in valence, arousal, and dominance dimensions are 6.8571, 5.8571, and 6.00 respectively.

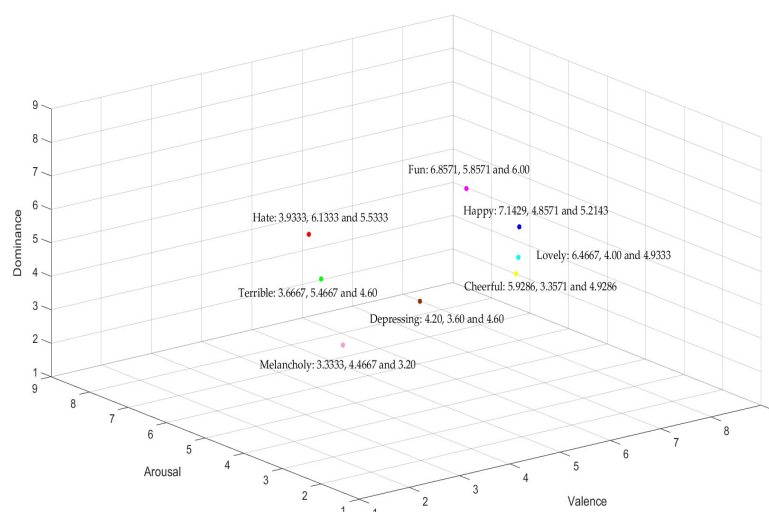


Figure 1. Three-dimensional space of the emotion model.

In the DEAP dataset, at the end of each trial, participants performed a self-assessment of their levels of arousal, valence, and dominance. Self-assessment manikins (SAMs) [37] were used to visualize the scales (see Figure 2). The manikins were displayed with the numbers 1–9 printed below. Participants were informed that they could click anywhere directly below or in-between the numbers, making the self-assessment a continuous scale [35]. In other words, the ratings given by each participant were real numbers within the interval 1–9. Therefore, thresholds need to be used to produce class labels. More categories can describe emotions more accurately. However, it greatly increases the difficulty of emotion recognition. In the preliminary stage of affective computing, many emotion recognition studies used label-processing methods based on self-assessment ratings of 1–5 and 5–9 [9,38–42].

Koelstra et al. [35] divided the task of recognizing emotion into different binary classification problems. Wang et al. [42] carried out a binary classification task in the two dimensions of arousal and valence.

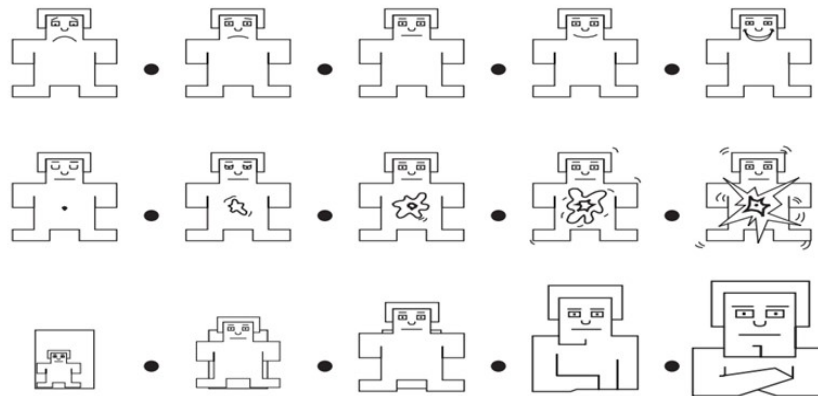


Figure 2. Images used for self-assessment: (top) Valence self-assessment manikin (SAM); (middle) Arousal SAM; (bottom) Dominance SAM.

There are three indicators: arousal, valence and dominance. Each trial was divided into two classes on each indicator. If the personal rating was less than 5, the label was set to “low”. If it was greater than or equal to 5, the label was set to “high” [9,38–42]. Thus, there were a total of six labels: HA (high arousal), LA (low arousal), HV (high valence), LV (low valence), HD (high dominance), and LD (low dominance) in three dimensions. Specifically, we divided the emotion recognition task into three binary classification problems.

3. Methods

3.1. Multiband Feature Matrix

The international 10–20 system describes the position of the scalp electrodes, and it is based on the relationship between electrode positions and the underlying region of the cerebral cortex. The “10” and “20” refer to the fact that the actual distances between adjacent electrodes are either 10% or 20% of the total front–back or the right–left distance of the skull. Figure 3 shows the specific form of the International 10–20 System and its mapping on a square matrix. The electrode positions marked with orange on the left side are the electrodes contained in the DEAP dataset. In this study, the 32 electrodes used in the DEAP dataset were mapped into a 9×9 square matrix. The orange points in the grid matrix represent the electrodes used in the international 10–20 system, and are filled with the frequency feature (PSD) of the EEG signals. The gray points are unused electrodes and are filled with default values. The nasion is represented by a gray triangle above the square matrix.

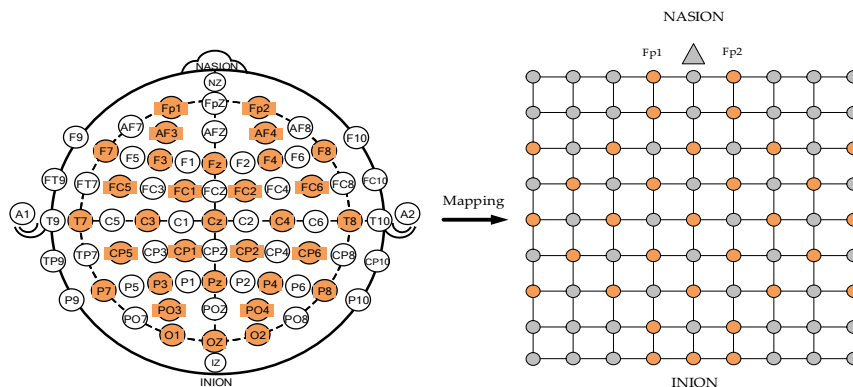


Figure 3. International 10–20 system and 9×9 square matrix.

Fusion of the power features from different frequency bands has been used in emotion recognition studies [4,6]. Therefore, multiple sub-matrices were constructed to describe the salient information related to the emotion states in different frequency bands, and each sub-matrix corresponded to a frequency band. Then, these sub-matrices were merged according to specific rules to structure the MFM. The MFM construction process is presented in Figure 4.

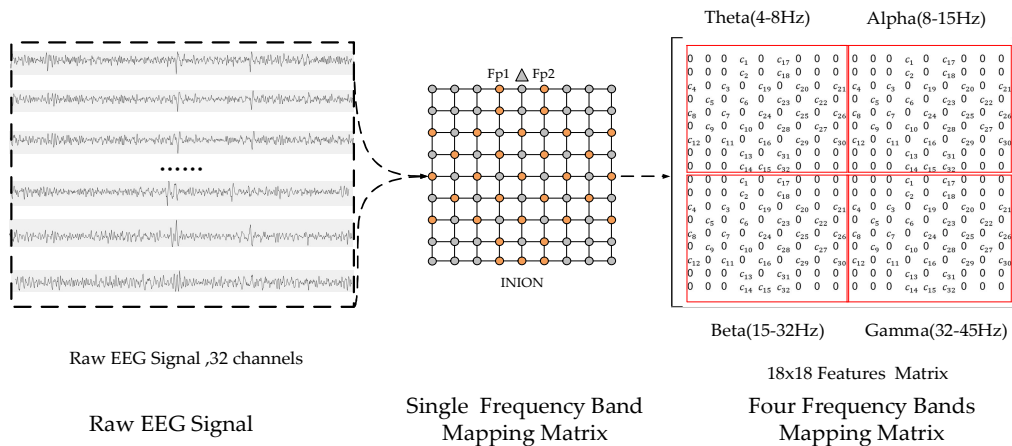


Figure 4. The mapping process of multiband feature matrix according to the raw electroencephalograph (EEG) signals of 32 channels.

The power spectral density (PSD) was extracted from the raw signal as a frequency domain feature. The PSD of each channel was divided into four parts according to the four frequency bands of theta (4–8 Hz), alpha (8–15 Hz), beta (15–32 Hz), and gamma (32–45 Hz). The average PSD value of each frequency band was taken as a frequency-domain feature value. Thus, four values were obtained for one channel. The number of features for each sample was 128 (32 channels and 4 frequency bands). The features of each subject were normalized using Equation (1), and the interval was from 0 to 1:

$$F' = \frac{F_{\max} - F}{F_{\max} - F_{\min}}, \tag{1}$$

where F' is the normalized feature value. F_{\max} and F_{\min} represent the maximum and minimum feature values, respectively. F is the feature value before normalization.

After normalization, each sample took the features of 32 channels with the same frequency band and constructed a sub-matrix according to the mapping rules described in Figure 3. The element values in the sub-matrix were set to the average PSD of the corresponding channels, and elements without corresponding electrodes in the sub-matrix had a default value of 0.

Considering that there are four bands (i.e., theta, alpha, beta, and gamma), four sub-matrices were obtained for each sample. The four sub-matrices were combined into an MFM in a 2×2 manner, as shown in Figure 4. Thus, an MFM of size 18×18 was constructed. In each sub-matrix, c_j represents the PSD value of the j -th channel in the corresponding frequency band.

In the MFM, the elements in each sub-matrix used to represent the EEG channels were consistent with the positions of sensors placed on the scalp when the EEG signals were acquired. The elements' values were the average PSD values of the corresponding channels. The MFMs showed the variations of human EEG signals on the scalp directly and accurately. Compared with traditional features, MFMs—which combine the frequency domain, spatial, and frequency band characteristics of multi-channel EEG signals—can provide richer information representing EEG signal variations in different emotion states.

3.2. Capsules Network

The CNN is the most commonly used method for two-dimensional object classification. However, information such as position and pose in the object is discarded by the CNN due to its data routing procedure. To compensate for the shortcomings of CNN, a network structure called the capsule network has been proposed [43]. CapsNet is a deep network model consisting of capsules. A capsule is composed of a group of neurons. Activation neurons represent the characteristics of components in the object. Each capsule is responsible for determining a single component in the object, and all capsules jointly determine the overall structure of the object. In contrast to some deep neural networks (e.g., DNN), this structure preserves objects' components and spatial information. Similar to CNN, a CapsNet is composed of a multi-layer network. Figure 5 shows the structure of a capsule and how information is routed from the lower-level capsules to the subsequent higher-level capsules.

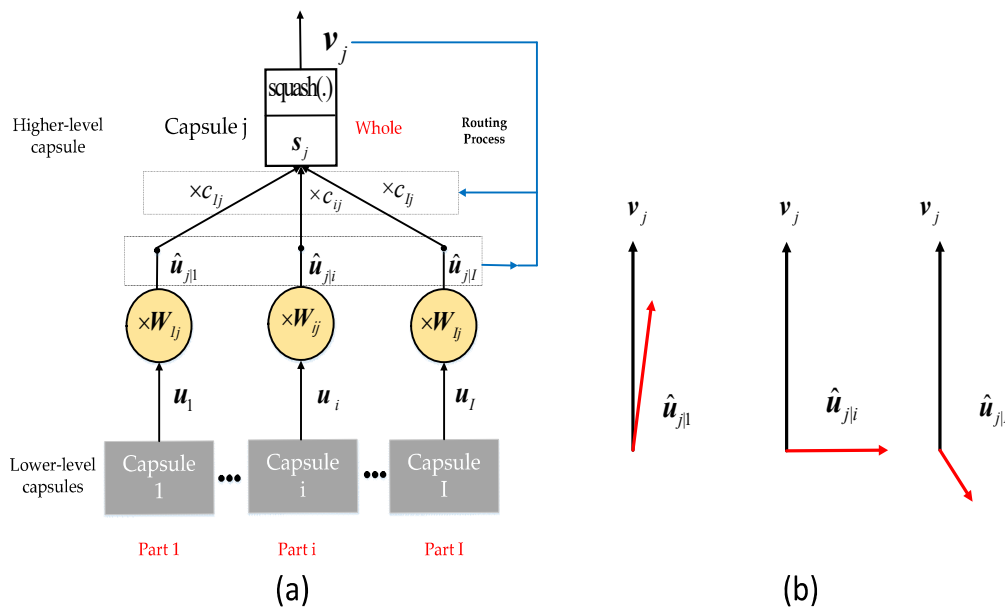


Figure 5. (a) Information transfer between capsules; (b) Routing process.

In part (a) of Figure 5, the inputs and outputs of the capsules are vectors. The length of the output u_i represents the probability of existence of its corresponding component, and the directions of the vector u_i encode various properties (e.g., size, position) of its corresponding component. The predictive vector \hat{u}_{ji} represents the belief, which encodes the relationship between the i -th capsule in the lower-level capsules and j -th capsule in the higher-level capsules using a linear transformation matrix W_{ij} in Equation (2):

$$\hat{u}_{ji} = W_{ij} \cdot u_i. \tag{2}$$

That is, the detected component's existence and pose information are used to predict the whole presence and pose information. During the training process, the network gradually learns to adjust the transformation matrix of the capsule pair through the corresponding relationship between components and the whole in the object.

In the higher-level capsules, s_j and v_j are the input and output of capsule j , respectively. s_j represents the sum values of the predicted vectors \hat{u}_{ji} with corresponding weight c_{ij} in low-level capsule i . In Equation (3), c_{ij} is the coupling coefficient and is determined by an iterative dynamic routing algorithm, where $\sum_j c_{ij} = 1$ and $c_{ij} \geq 0$. When $c_{ij} = 0$, there is no information transfer between capsule i and capsule j , whereas when $c_{ij} = 1$, all the information of capsule i is transmitted to the high-level capsule j . Since the length of the output represents a probability value, a non-linear squash

function is used to ensure that the short vector is compressed to be closer to 0 and the long vector is compressed to be closer to 1. The squash function is shown in Equation (4):

$$\mathbf{s}_j = \sum_i c_{ij} \cdot \hat{\mathbf{u}}_{j|i}, \quad (3)$$

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j^2\|}{1 + \|\mathbf{s}_j^2\|} \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|}, \quad (4)$$

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}, b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i} \cdot \mathbf{v}_j. \quad (5)$$

As shown in part (b) of Figure 5, when the lower-level capsules and the higher-level capsules are consistent with their predictions, the value of c_{ij} becomes larger by Equation (5), and it gets smaller when they are inconsistent. By adjusting the routing coefficient, the dynamic routing algorithm ensures that the lower-level capsules send their prediction vectors to the higher-level capsules that are consistent with their predictions, so that the outputs of the sub-capsules are sent to the correct parent capsules. The specific process of the dynamic routing algorithm can be found in [43].

3.3. Capsule-Network-Based Deep Learning Model

A CapsNet can encode the relative relationship between local parts and whole objects (e.g., scale, position, direction), which is discarded by CNNs in the process of max-pooling and average-pooling. Therefore, CapsNet has a natural understanding of three-dimensional space and contains the states and relative positional relationships between all components of an object. Thus, it is more sensitive to differences in spatial information between different objects. The components in an image are represented by the values of the pixels in the corresponding region. Similarly, in different emotional states, different regions of the cerebral cortex have corresponding responses. Corresponding to the MFM, this is the features of EEG channels in different regions. In view of this, CapsNet can distinguish the salient global difference information reflected by the human brain in different emotional states when it is applied to multi-channel EEG-based emotion recognition. In addition, CapsNet has a faster learning speed and less sample usage than CNN. Therefore, CapsNet is introduced into EEG-based multi-channel emotion recognition, and a deep learning framework based on an MFM and a capsule network is proposed.

The CapsNet-based model consists of four parts. The first part is convolution with rectified linear unit (ReLU), and convolution operations are performed on the input MFM to detect local features in this part. Primary capsules (PrimaryCaps) is the second part, which consists of a convolution process and transmits information to capsules. The third part is Emotion capsules (EmotionCaps), which includes the dynamic routing process between capsules and is used for emotion recognition. The last part attempts to reconstruct the input matrix from the output of the final capsules.

Figure 6 describes the structure of the proposed CapsNet-based model. In the convolution with ReLU, an 18×18 MFM which is described in Section 3.1 is decoded as the input in the first layer. The first step is the standard convolution operation; a 3×3 convolution kernel with a stride size of 1 and a ReLU activation function is employed. This convolutional layer does not use padding, and the output volume is 16×16 with 256 channels.

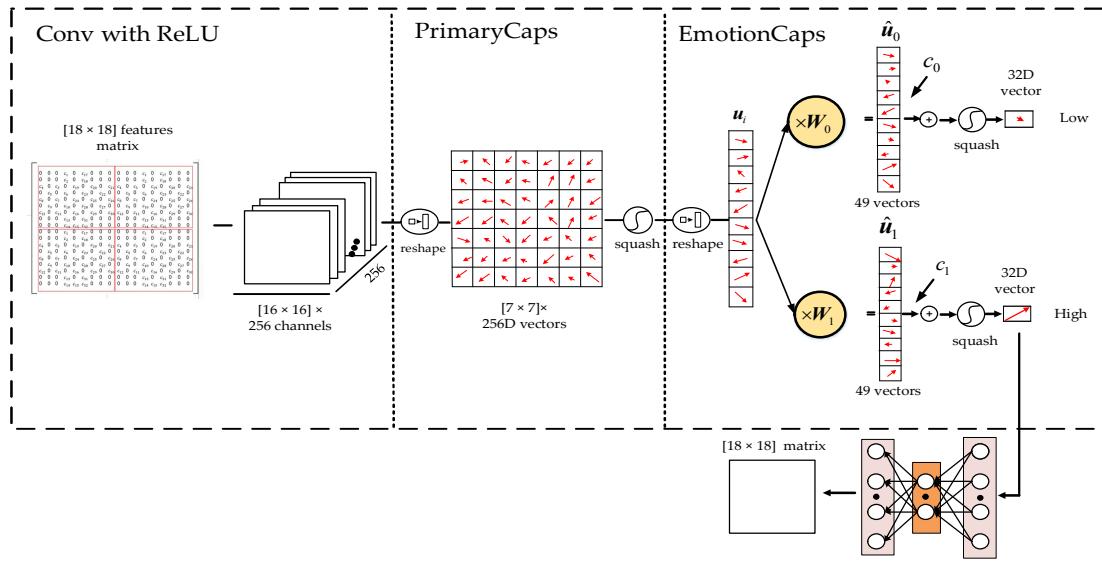


Figure 6. Architecture of the capsule network (CapsNet)-based model. ReLU: rectified linear unit.

In the PrimaryCaps part, convolution operations are first performed on the data from the previous part, and the convolution layers are composed of a convolution unit with a convolution kernel size of 3×3 and a stride of 2. After the convolution operation, the output data are reshaped to 256-dimensional (256D) vectors in 7×7 arrays. Then, the squash activation function is applied.

The EmotionCaps part reshapes the 7×7 arrays generated from the PrimaryCaps, and forms 49 by the 256D vector. Then, the vector is multiplied by weight matrix W_i , and the vector \hat{u}_i can be obtained, where i denotes the index of each output class. Then, c_i is determined by the dynamic routing algorithm. The dynamic routing algorithm iterates three times in this model, and c_i encodes the \hat{u}_i into a 32-dimensional activation vector of instantiation parameters. Lastly, the output vector is squashed in order to determine the probability of each emotion state. The values of i are 0 and 1, which correspond to the respective emotion classes.

The final part attempts to reconstruct the input MFM from the final capsules, which keeps the information from the input as much as possible throughout the network. Thus, it can prevent overfitting and help to generalize new samples as a normalizer. The structure uses a three-layer feed-forward neural network with 512, 1024, and 324 units, respectively. This model uses L2 for reconstruction and margin loss for classification. The margin loss is shown in Equation (6):

$$L_e = T_e \max(0, m^+ - \|v_e\|)^2 + \lambda(1 - T_e) \max(0, \|v_e\| - m^-)^2, \quad (6)$$

where T_e equals 1 if an emotional class e is present and $m^- = 0.1$ and $m^+ = 0.9$. The λ is down-weighting of the loss and it was set to 0.5 by default. v_e represents the final output vector of class e .

4. Results and Discussions

In order to increase the number of samples, the raw EEG signals of each channel were divided into 20 sections by using a sliding window. The duration of the sliding window was 3 s, and there was no overlap region between adjacent windows. Each section was regarded as an independent sample, and inherited the labels of the original sample. The number of EEG samples per subject was 800 (40×20), and the number of all samples was 25,600 (800×32). The total number of MFMs of all subjects was 25,600, corresponding to the number of samples, as shown in Table 1.

Table 1. Distribution of samples. HA: high arousal; HD: high dominance; HV: high valence; LA: low arousal; LD: low dominance; LV: low valence.

Label	Data Quantity	Label	Data Quantity	Label	Data Quantity
HA	14,740	HV	14,160	HD	15,600
LA	10,860	LV	11,440	LD	10,000
Total	25,600	Total	25,600	Total	25,600

4.1. Selection of Hyperparameters

Each sample constructed a corresponding MFM. Since MFMs of single subjects are not sufficient for training a stable recognition model, the MFMs of all subjects were used to improve the generalization ability of the proposed method.

The classification performance was analyzed via a 10-fold cross validation technique. For this work, the 25,600 MFMs were divided into 10 subsets. Nine subsets were assigned to the training set, and the remaining one was assigned to the test set. The above process was repeated 10 times until all subsets were tested. The training epochs and batch size for each experiment were set to 400 and 40, respectively. Dynamic routing algorithms in all models were iterated three times.

Different combinations of hyperparameters and training parameters were tested to determine the appropriate parameter combination of the proposed model, and the parameter combination with the minimal recognition error was adopted. Considering that the computational complexity of the proposed model will increase sharply as the number of layers increases, the hyperparameters included only the parameters of the first convolution layer, filters of PrimaryCaps, and parameters of EmotionCaps. The number of layers of the CapsNet remained unchanged.

The number and form of trainable parameters affect the performance of the capsule network. To find the optimal combination of model parameters, a two-pass search process from coarse to fine was adopted. Firstly, we made a coarse selection to determine the scope of the number of model parameters. The network parameters were selected by referring to references [43,44]. Table 2 shows the network models with different numbers of parameters in the coarse selection step.

Table 2. Five CapsNet-based models with different parameters in the first step.

Input Data	Model	Conv. with ReLU	Primary Capsule	EmotionCaps	CapsNet Output
$(18 \times 18) \times 1$	1	$3 \times 3 \times 32$	$3 \times 3 \times (1 \times 32)$	$(7 \times 7) \times 32$	2×16
	2	$3 \times 3 \times 128$	$3 \times 3 \times (1 \times 128)$	$(7 \times 7) \times 128$	2×16
	3	$3 \times 3 \times 256$	$3 \times 3 \times (1 \times 256)$	$(7 \times 7) \times 256$	2×16
	4	$3 \times 3 \times 384$	$3 \times 3 \times (1 \times 384)$	$(7 \times 7) \times 384$	2×16
	5	$3 \times 3 \times 512$	$3 \times 3 \times (1 \times 512)$	$(7 \times 7) \times 512$	2×16

Figure 7 shows the emotion recognition results of the models in the coarse selection step. Compared with model 1, model 2 and model 3 achieved better recognition performance in three dimensions. The reason may be that model 1 had fewer parameters, so it could not fully learn the features of MFMs. The four remaining models had similar recognition performance. However, model 3 and model 4 had higher computational complexity. Therefore, model 2 and model 3 were analyzed further.

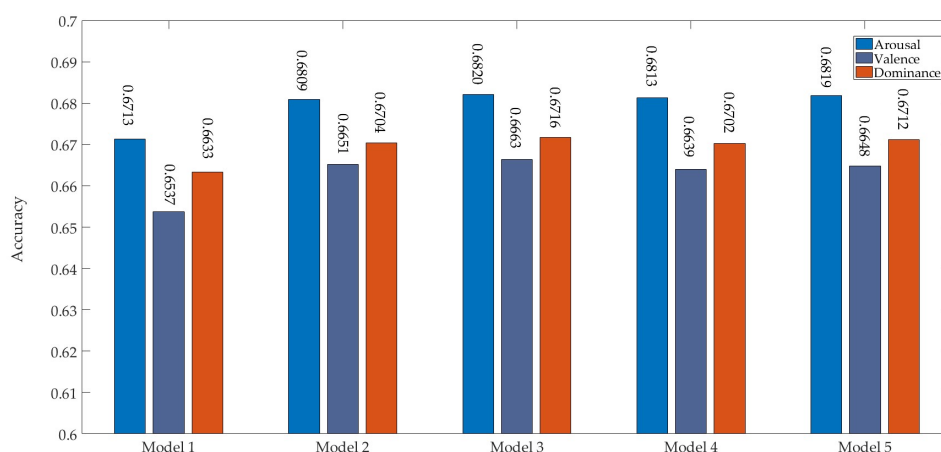


Figure 7. Accuracy of different models in three dimensions in the first step.

When the scope of the number of parameters was determined, several models which had similar numbers of parameters and different network structures were employed to determine the optimal parameter combination. Specifically, we adjusted the parameters of the three main parts of model 2 and model 3, including the size of the convolution kernel in the first part, the parameters of the Primary part, and the parameters of the EmotionCaps part. Table 3 shows the combination of parameters used by different network models.

Table 3. Six CapsNet-based models (A, B, C, D, E, and F) with different parameters in the second step.

Input Data	Model	Conv with ReLU	Primary Capsule	EmotionCaps	CapsNet Output
$(18 \times 18) \times 1$	A	$3 \times 3 \times 256$	$3 \times 3 \times (1 \times 256)$	$(7 \times 7) \times 256$	2×32
	B	$5 \times 5 \times 256$	$3 \times 3 \times (1 \times 256)$	$(6 \times 6) \times 256$	2×16
	C	$9 \times 9 \times 256$	$3 \times 3 \times (32 \times 8)$	$(4 \times 4 \times 32) \times 8$	2×32
	D	$3 \times 3 \times 128$	$3 \times 3 \times (1 \times 128)$	$(7 \times 7) \times 128$	2×32
	E	$5 \times 5 \times 128$	$3 \times 3 \times (1 \times 128)$	$(6 \times 6) \times 128$	2×16
	F	$9 \times 9 \times 128$	$3 \times 3 \times (16 \times 8)$	$(4 \times 4 \times 16) \times 8$	2×32

Figure 8 shows the emotion recognition results of the evaluation method by using different models in the second step. Model A achieved the highest recognition accuracies in all three dimensions (i.e., 0.6828, 0.6673, and 0.6725 in arousal, valence, and dominance, respectively). The reason may be that the output volumes from the first convolution operation of the model A were larger, which is more beneficial for preserving the features of the matrix. Furthermore, the vectors of model A in EmotionCaps layer had higher dimension, and also contained more spatial information. The training speeds of models A, B, D, and E were more than twice those of models C and F. Thus, the hyperparameters of model A were adopted. It can be seen from the comparison of the models that the output volume of the first convolutional layer and the vectors of the PrimaryCaps had an influence on the classification performance.

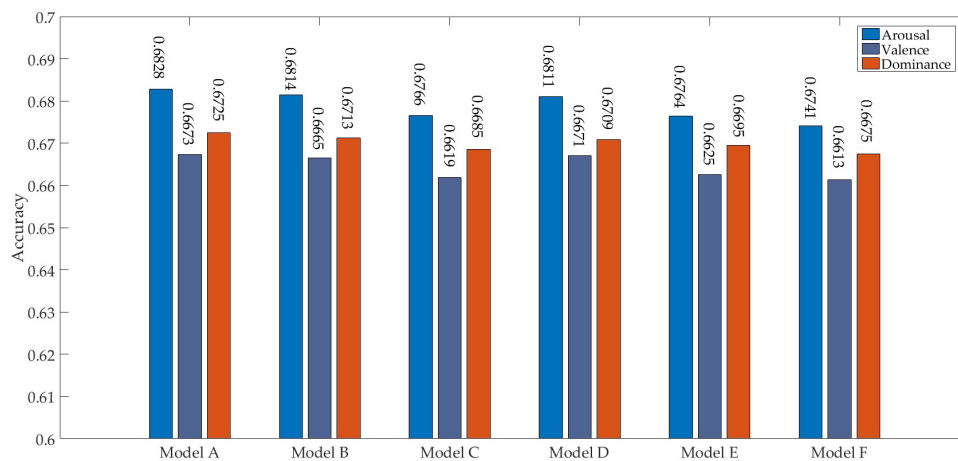


Figure 8. Accuracy of different models in three dimensions in the second step.

4.2. Comparison between the CapsNet Model and Common Classifiers

To verify the effectiveness of the CapsNet-based model in utilizing frequency-domain, spatial-domain, and frequency-band characteristics of multi-channel EEG signals for emotion recognition, we also compared the average recognition performance between the CapsNet-based model and several common models which also use MFM as the input feature (Figure 9). These models included 1D-CNN, 2D-CNN, random decision forest (RDF), k-nearest neighbor, and support vector machines. All methods used a 10-fold cross validation technique, and 25,600 MFMs were employed.

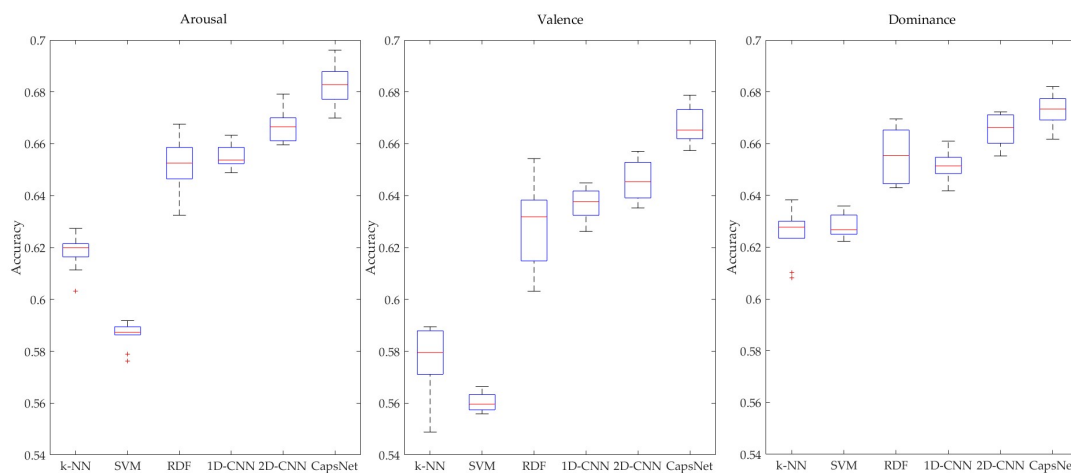


Figure 9. Comparison of emotion recognition results between CapsNet and comparison classifiers. CNN: convolutional neural network; k-NN: k-nearest neighbor; RDF: random decision forest; SVM: support vector machine.

Specifically, 1D-CNN and 2D-CNN were implemented by TensorFlow and the structures were similar to the capsule network constructed in the proposed model. In the 1D-CNN model, the input data were 128-dimensional vectors (32 channels, 4 frequency bands) that were not rearranged. The first convolution layer employed a 1×9 kernel with a stride size of 1 and a ReLU activation function. The number of filters in this layer was 256. The second convolutional layer had the same parameters as the first convolutional layer. The third layer also had the same convolutional parameters as the first layer, except that the number of filters was 128. Each convolutional layer corresponded to a max pooling layer, which implemented a max pooling process for information aggregation after the convolution operation was completed. The parameters of the three max pooling layers were the same. Each max pooling layer used a 1×2 kernel with a stride size of 1 and its padding was “SAME”. The node

numbers of the two full-connection layers were 324 and 162, respectively, and the output layer used a softmax function. In the 2D-CNN model, the convolution layer employed a 5×5 kernel and the max pooling layer used a 3×3 kernel. All parameters except those mentioned above were the same as for the 1D-CNN.

Other common models include RDF, which is based on ensemble learning theory; k-NN, which is based on distance metric ($neighbors = 5$); SVM, based on statistical learning theories ($c = 10$; $kernel = linear$). The above three models were implemented by the Scikit-learn toolkit [45]. In order to adapt to the data entry forms of the above classifiers, each row in the matrix was extracted to form a vector, and all vectors were connected to form a high-dimensional vector.

The recognition results of the CapsNet model A and the five other common models are illustrated in box plots in Figure 9. The average recognition accuracies of all subjects using CapsNet model A in the three dimensions of arousal, valence, and dominance were 0.6828, 0.6673, and 0.6725, respectively, whereas the average accuracies of other models were 0.6667, 0.6460, and 0.6658 with 2D-CNN; 0.6550, 0.6370, and 0.6513 with 1D-CNN; 0.6512, 0.6288, and 0.6547 with RDF; 0.5862, 0.5603, and 0.6286 with SVM; and 0.6182, 0.5770, and 0.6254 with k-NN. The recognition accuracies of 2D-CNN in three dimensions were higher than that of 1D-CNN. It can be seen that the rearranged features (MFMs) could provide additional spatial information representing EEG signal variations in different emotion states. The CapsNet model had the highest average accuracy. In addition, the difference between all experimental results of this method was relatively smaller than that of other methods. The reason may be that the proposed model based on a capsule network can make better use of the three features contained in the MFM and has stronger generalization ability. Compared with CapsNet, 1D-CNN and 2D-CNN could not retain salient spatial information between all channels and bands. Therefore, the classification accuracies of the two CNN models were lower. Due to the architecture and fundamentals of RDF, k-NN, and SVM, the MFM is converted into a high-dimensional vector. Therefore, they can only distinguish the difference of frequency-domain features in the same channels and cannot mine the salient spatial information between channels in different frequency bands, so their recognition performance was poor. We also applied a Wilcoxon signed-rank test ($\partial < 0.05$) to examine the performance between CapsNet and 2D-CNN in three dimensions. The null hypothesis was that the performance was similar. If the p -value was larger than ∂ , the null hypothesis was accepted. The p -values of arousal, valence, and dominance were 0.002, 0.002, and 0.0488, respectively. The results of the Wilcoxon signed-rank test showed that the performance of CapsNet was superior to 2D-CNN, with p -values less than 0.05 in three dimensions. The results show that spatial differences between channels in different frequency bands could provide salient information regarding emotion states and the model based on capsules network proposed in this paper could utilize the above information effectively.

In view of the similarity between 2D-CNN and CapsNet, hold-out validation was used to evaluate the two models for further comparison. The 25,600 MFMs were randomly divided into training, validation, and test sets, with respective proportions of 8:1:1 for each label. The results in the validation set are shown in Figure 10. With the same data, the validation accuracy of CapsNet increased faster in three dimensions, and the accuracy of CapsNet was higher than that of 2D-CNN with a similar network structure using the same epoch.

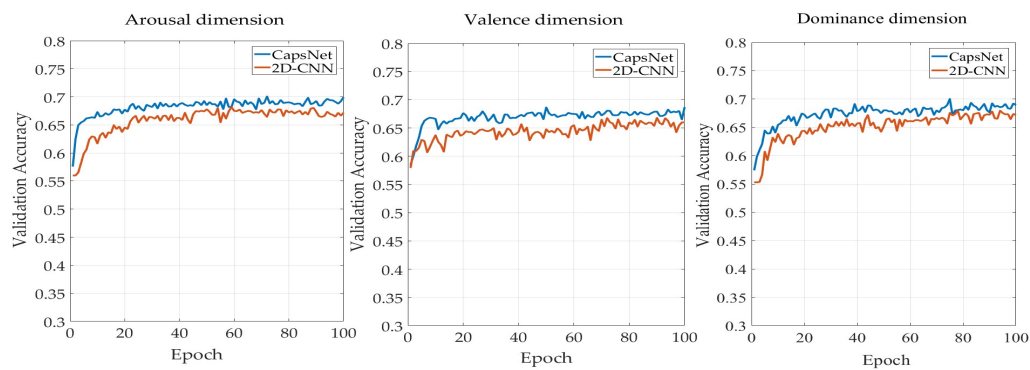


Figure 10. Comparison of 2D-CNN and CapsNet using multiband feature matrices (MFMs).

4.3. Comparison between the Proposed Methods and Several Existing Studies

Finally, we compared the proposed method with other methods using the same dataset. Table 4 shows the details of the comparison methods. Among them, dual-tree complex wavelet packet transform (DT-CWPT) time-frequency features were used in references [38,39]. Items crossed in the table are not indicated in the corresponding references.

Table 4. Details of several reported studies. DT-CWPT: dual-tree complex wavelet packet transform.

Studies	Features	Classifiers	Hyperparameters	Evaluation Methods	Accuracy		
					Arousal	Valence	Dominance
Koelstra et al. [35]	Power spectral features	Gaussian naive Bayes	–	leave-one-trial-out validation	0.6200	0.5760	–
Naser and Saha [38]	DT-CWPT	SVM	radial-basis kernel	leave-one-trial-out validation	0.6690	0.6530	0.691
Naser and Saha [39]	DT-CWPT	SVM	radial-basis kernel	leave-one-trial-out validation	0.6620	0.6430	0.689
Chung and Yoon [40]	Power spectral features	Bayes classifier	weighted-log-posterior function	10-fold cross-validation	0.6640	0.6660	–
Li et al. [9]	DBN features	SVM	radial-basis kernel	10-fold cross-validation	0.6420	0.5840	0.658
Wang and Shang [41]	DBN features	Deep belief networks	DBN structure: 512–50–50–2 learning rate: 0.05 epochs: 20 Mini-batch: 100	leave-one-trial-out validation	0.5120	0.6090	–
Zhuang et al. [42]	The first difference of time series, etc.	SVM	radial-basis kernel	leave-one-trial-out validation	0.6910	0.7199	–
The proposed method	Multiband Feature Matrix	CapsNet	hyperparameters of the model A	10-fold cross-validation	0.6828	0.6673	0.6725

As shown in Table 4, the proposed method achieved the highest recognition accuracy in arousal and valence dimensions, with the exception of the results of reference [42]. It can be seen that the proposed method could improve the performance of multi-channel EEG-based emotion recognition. The recognition accuracy in valence and arousal dimensions of reference [42] was higher than that in this work. The reason may be that the method in reference [42] constructs a subject-related classification model, which was only employed to recognize the samples belonging to the corresponding subject. Although the recognition accuracy of references [38] and [39] in the dominance dimension was higher than that of our work, it was lower in the other two dimensions than our method. In order to improve the generalization of the model, the proposed framework used all the data of all subjects for emotion recognition, without considering the differences between the subjects. The results showed that the proposed framework also achieved good performance in all subjects' EEG-based emotion recognition task. Regardless, the framework based on MFM and CapsNet achieved satisfactory results for EEG emotion recognition.

5. Conclusions

In this paper, a framework using multi-channel EEG signals and a deep learning model was proposed to recognize human emotions. The innovation of this work is in two main aspects. Firstly, a novel feature extraction is proposed. Specifically, the frequency-domain features are extracted from EEG signals, and are mapped to an MFM according to electrode positions and frequency band information. The MFM reflects not only the relative position relationship between EEG channels, but also the difference of frequency bands in the EEG frequency domain features. Thus, it provides salient information related to emotion states in spatial and frequency domains. Meanwhile, we also propose a deep learning model based on a capsule network, which was used to judge emotional state. This is the first attempt to apply a capsule network to multi-channel EEG emotion recognition.

The experimental results show that the proposed method achieved satisfactory results in two indicators. This demonstrates that the three kinds of information contained in the proposed MFM features are helpful to recognize human emotion, and the model based on CapsNet could effectively use this information for emotion recognition. At present, our method has only been tested on DEAP datasets. We will test it on more emotional datasets to verify the method comprehensively in the future.

Author Contributions: Conceptualization, H.C.; methodology, H.C. and L.D.; software, L.D.; validation, L.D. and Y.L.; formal analysis, H.C.; investigation, B.L.; resources, B.L.; data curation, Y.L.; writing—original draft preparation, H.C. and L.D.; writing—review and editing, Y.L.; visualization, B.L.; supervision, B.L.; project administration, H.C.; funding acquisition, H.C.

Funding: This research was funded by the National Natural Science Foundation of China under Grants No. 61502150 and No. 61403128, Foundation for University Key Teacher by Henan Province under Grant No. 2015GGJS068, Fundamental Research Funds for the Universities of Henan Province under Grant No. NSFRF1616, Foundation for Scientific and Technological Project of Henan Province under Grant No. 172102210279, and Key Scientific Research Projects of Universities in Henan Grant No. 19A520004.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Adolphs, R.; Tranel, D.; Damasio, H.; Damasio, A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **1994**, *372*, 669–672. [[CrossRef](#)] [[PubMed](#)]
2. Chatterjee, M.; Zion, D.J.; Deroche, M.L.; Burianek, B.A.; Limb, C.J.; Goren, A.P.; Kulkarni, A.M.; Christensen, J.A. Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Hearing Res.* **2015**, *322*, 151–162. [[CrossRef](#)] [[PubMed](#)]
3. Ross, P.D.; Polson, L.; Grosbras, M.H. Developmental Changes in Emotion Recognition from Full-Light and Point-Light Displays of Body Movement. *Plos One* **2012**. [[CrossRef](#)]
4. Zhong, Y.; Zhao, M.; Wang, Y.; Yang, J.; Zhan, J. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Comput. Meth. Prog. Bio.* **2017**, *140*, 93–110.
5. Abadi, M.K.; Kia, M.; Subramanian, R.; Avesani, P.; Sebe, N. Decoding affect in videos employing the MEG brain signal. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–6.
6. Chao, H.; Zhi, H.Z.; Dong, L.; Liu, Y.L. Recognition of Emotions Using Multichannel EEG Data and DBN-GC-Based Ensemble Deep Learning Framework. *Comput. Intel. Neurosc.* **2018**, *2018*, 9750904. [[CrossRef](#)]
7. Li, Y.; Huang, J.; Zhou, H.; Zhou, H.Y.; Zhong, N. Human Emotion Recognition with Electroencephalographic Multidimensional Features by Hybrid Deep Neural Networks. *Appl. Sci.* **2017**, *7*, 1060. [[CrossRef](#)]
8. Zheng, W.L.; Dong, B.N.; Lu, B.L. Multimodal emotion recognition using EEG and eye tracking data. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014.
9. Li, X.; Zhang, P.; Song, D.W.; Yu, G.L.; Hou, Y.X.; Hu, B. EEG Based Emotion Identification Using Unsupervised Deep Feature Learning. In Proceedings of the SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research, Santiago, Chile, 13 August 2015.

10. Bilalpur, M.; Kia, S.M.; Chawla, M.; Chua, T.-S.; Subramanian, R. Gender and emotion recognition with implicit user signals. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, Scotland, 13–17 November 2017.
11. Zheng, W.L.; Zhu, J.Y.; Yong, P.; Lu, B.L. EEG-based emotion classification using deep belief networks. In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 14–18 July 2014.
12. Liu, W.; Zheng, W.L.; Lu, B.L. Multimodal Emotion Recognition Using Multimodal Deep Learning. Available online: <https://arxiv.org/abs/1602.08225> (accessed on 30 September 2016).
13. Liu, Y.; Sourina, O. Real-Time Fractal-Based Valence Level Recognition from EEG. *IEEE Trans. Serv. Comput.* **2013**, *7848*, 101–120.
14. Murugappan, M.; Ramachandran, N.; Sazali, Y. Classification of human emotion from EEG using discrete wavelet transform. *JBiSE*. **2010**, *3*, 390–396. [[CrossRef](#)]
15. Chai, T.Y.; San, W.S.; Tan, C.S.; Rizon, M. Classification of human emotions from EEG signals using statistical features and neural network. *Int. J. Eng. Sci.* **2009**, *1*, 1–6.
16. Frantzidis, C.; Bratsas, C.; Papadelis, C.; Konstantinidis, E.; Pappas, C.; Bamidis, P.D. Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *4*, 589–597. [[CrossRef](#)]
17. Hjorth, B. EEG analysis based on time domain properties. *Electroencephalogr. Clin. Neurophysiol.* **1970**, *29*, 306–310. [[CrossRef](#)]
18. Kroupi, E.; Yazdani, A.; Ebrahimi, T. EEG correlates of different emotional states elicited during watching music videos. In Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction, Memphis, TN, USA, 9–12 October 2011; Springer: Berlin/Heidelberg, Germany; pp. 457–466.
19. Petrantonakis, P.C.; Hadjileontiadis, L.J. Emotion recognition from EEG using higher order crossings. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 186–197. [[CrossRef](#)] [[PubMed](#)]
20. Petrantonakis, P.C.; Hadjileontiadis, L.J. Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *IEEE Trans. Affect. Comput.* **2010**, *1*, 81–97. [[CrossRef](#)]
21. Jenke, R.; Peer, A.; Buss, M. Feature Extraction and Selection for Emotion Recognition from EEG. *IEEE Trans. Affect. Comput.* **2017**, *5*, 327–339. [[CrossRef](#)]
22. Yin, Z.; Wang, Y.; Liu, L.; Zhang, W.; Zhang, J. Cross-Subject EEG Feature Selection for Emotion Recognition Using Transfer Recursive Feature Elimination. *Front. Neurobot.* **2017**, *11*, 19. [[CrossRef](#)]
23. 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. [[CrossRef](#)]
24. Lin, Y.-P.; Wang, C.-H.; Jung, T.P.; Wu, T.L.; Jeng, S.K.; Duann, J.R.; Chen, J.H. EEG-based emotion recognition in music listening. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 1798–1806.
25. Liu, Y.J.; Yu, M.; Zhao, G.; Song, J.; Ge, Y.; Shi, Y. Real-time movie-induced discrete emotion recognition from EEG signals. *IEEE Trans. Affect. Comput.* **2017**, *9*, 550–562. [[CrossRef](#)]
26. Thammasan, N.; Moriyama, K.; Fukui, K.; Numao, M. Continuous music-emotion recognition based on electroencephalogram. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1234–1241. [[CrossRef](#)]
27. Rozgic, V.; Vitaladevuni, S.N.; Prasad, R. Robust EEG emotion classification using segment level decision fusion. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.
28. Wichakam, I.; Vateekul, P. An evaluation of feature extraction in EEG-based emotion prediction with support vector machines. In Proceedings of the 2014 11th international Joint Conference on Computer Science and Software Engineering, Chon Buri, Thailand, 14–16 May 2014.
29. Jie, X.; Cao, R.; Li, L. Emotion recognition based on the sample entropy of EEG. *Bio-med. Mater. Eng.* **2014**, *24*, 1185.
30. Li, M.; Xu, H.P.; Liu, X.W.; Lu, S.F. Emotion recognition from multichannel EEG signals using K-nearest neighbor classification. *Technol. Health Care* **2018**, *26*, 509–519. [[CrossRef](#)]
31. Sohaib, A.T.; Qureshi, S.; Hagelbäck, J.; Hilborn, O.; Jerica, P. Evaluating classifiers for emotion recognition using EEG. In Proceedings of the International conference on augmented cognition, Berlin, Heidelberg, 21–26 July 2013; pp. 492–501.
32. Jirayucharoensak, S.; Panngum, S.; Israsena, P. EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World J.* **2014**, *2014*, 627892. [[CrossRef](#)]

33. Liu, W.; Zheng, W.L.; Lu, B.L. Emotion Recognition Using Multimodal Deep Learning. In Proceedings of the 23rd International Conference on Neural Information Processing, Kyoto, Japan, 16–21 October 2016; pp. 521–529.
34. Yang, B.; Han, X.; Tang, J. Three class emotions recognition based on deep learning using staked autoencoder. In Proceedings of the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 14–16 October 2017.
35. Koelstra, S.; Muhl, C.; Soleymani, M.; Jong-Seok, L.; 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. [[CrossRef](#)]
36. Verma, G.K.; Tiwary, U.S. Affect representation and recognition in 3D continuous valence–arousal–dominance space. *Multimed. Tools Appl.* **2017**, *76*, 2159–2183. [[CrossRef](#)]
37. Morris, J.D. SAM: The Self-Assessment Manikin an Efficient Cross-Cultural Measurement of Emotional Response. *Advertising Research.* **1995**, *35*, 63–68.
38. Naser, D.S.; Saha, G. Classification of emotions induced by music videos and correlation with participants' rating. *Expert Syst. Appli.* **2014**, *41*, 6057–6065.
39. Naser, D.S.; Saha, G. Recognition of emotions induced by music videos using DT-CWPT. In Proceedings of the IEEE Indian Conference on Medical Informatics and Telemedicine (ICMIT), Kharagpur, India, 28–30 March 2013; pp. 53–57.
40. Chung, S.Y.; Yoon, H.J. Affective classification using Bayesian classifier and supervised learning. In Proceedings of the 2012 12th International Conference on Control, Automation and Systems, Je Ju Island, Korea, 17–21 October 2012; pp. 1768–1771.
41. Wang, D.; Shang, Y. Modeling Physiological Data with Deep Belief Networks. *Int. J. Inf. Educ. Technol.* **2013**, *3*, 505–511.
42. Zhuang, N.; Zeng, Y.; Tong, L.; Zhang, C.; Zhang, H.; Yan, B. Emotion Recognition from EEG Signals Using Multidimensional Information in EMD Domain. *BioMed Res. Int.* **2017**, *2017*. [[CrossRef](#)]
43. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 3859–3869.
44. Mobiny, A.; Van, N.H. Fast CapsNet for Lung Cancer Screening. Available online: <https://arxiv.org/abs/1806.07416> (accessed on 19 June 2018).
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2012**, *12*, 2825–2830.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).