Cognitive Computation
https://doi.org/10.1007/s12559-017-9533-x

Hierarchical Convolutional Neural Networks for EEG-Based Emotion Recognition

Jinpeng Li[1,2] · Zhaoxiang Zhang[1,2,3] · Huiguang He[1,2,3]

Received: 10 April 2017 / Accepted: 30 November 2017

Abstract
Traditional machine learning methods suffer from severe overfitting in EEG-based emotion reading. In this paper, we use hierarchical convolutional neural network (HCNN) to classify the positive, neutral, and negative emotion states. We organize differential entropy features from different channels as two-dimensional maps to train the HCNNs. This approach maintains information in the spatial topology of electrodes. We use stacked autoencoder (SAE), SVM, and KNN as competing methods. HCNN yields the highest accuracy, and SAE is slightly inferior. Both of them show absolute advantage over traditional shallow models including SVM and KNN. We confirm that the high-frequency wave bands Beta and Gamma are the most suitable bands for emotion reading. We visualize the hidden layers of HCNNs to investigate the feature transformation flow along the hierarchical structure. Benefiting from the strong representational learning capacity in the two-dimensional space, HCNN is efficient in emotion recognition especially on Beta and Gamma waves.

Keywords Affective brain-computer interface · Emotion recognition · Brain wave · Deep learning · EEG

Introduction

Emotion is important when people interact with each other. In some cases, we also want the machines to interact with us according to our emotion states. The affective brain-computer interface (aBCI) [1] is a technique to make it a reality. In aBCIs, the key problem is to deduce their human user's emotion state, which is the foundation of other functional modules, e.g., emotion feedback. Engineers hold strong interest in emotion recognition techniques as they are valuable in multiple applications, e.g., workload estimation [2], driving fatigue detection [3], and mental state monitoring for pilots [4]. At the same time, the computational models of emotion might also help psychologists understand the internal mechanism of human emotion processing.

There are many clues that contain emotion information. The first kind is called the "non-physiological clue," e.g., facial expression and gesture [5, 6]. For example, [7] fused expression, speech, and other multimodal information together and discriminated four emotion states. The non-physiological clues are relatively economical. However, signals of this kind are highly related to the personal habit (culture) of subjects, and thus could not be universally applied. Several studies have been conducted concerning this topic [8–10]. The second kind is called the "physiological clue," e.g., the electric activities of neural cell clusters across the human cerebral cortex. EEG is used to record such activities. EEG is reliable in emotion recognition, because it has relatively objective evaluation on emotion in comparison with the non-physiological ones [5]. EEG devices often have multiple channels (electrodes) to collect electric potentials from different positions, which are either implantable or non-implantable. The former has relatively higher signal-to-noise ratio (SNR), but not fit to the daily use. The latter is noninvasive and wearable, but the SNR is lower, which is a challenge to classification. For the convenience in practice, the non-implantable EEG is favorable to commercial aBCI [10].

✉ Huiguang He
huiguang.he@ia.ac.cn

Jinpeng Li
lijinpeng2015@ia.ac.cn

Zhaoxiang Zhang
zhaoxiang.zhang@ia.ac.cn

[1] Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2] University of Chinese Academy of Sciences (UCAS), Beijing, China

[3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China

Published online: 16 December 2017

EEG signals are time-domain series, and we could analyze them either in the time domain [11] or in the frequency domain [12], or the combination of them. There are five bands of brain waves that interest researchers: Delta, Theta, Alpha, Beta, and Gamma. Delta waves (1–3Hz) [13] are the slowest "sleep waves," which are often used to characterize the depth of sleep. Theta waves (4–7 Hz) [14] are believed to be active in light meditation and sleeping. Alpha waves (8–13 Hz) [15] are linked with relaxation, happiness, and well-being. The former three bands are associated with sleep or relaxation, but the remaining two bands are different. Beta waves (14–30 Hz) [16] are the "waking consciousness and reasoning waves," which are tightly related to our thinking process. Beta waves are significant in the effective functioning throughout the day. Researchers proved that they may translate into stress, anxiety, and restlessness. This band has been intensively investigated. Gamma waves (31–50 Hz) [17] have the highest frequency, and we know little about them. Initial research shows Gamma waves are associated with bursts of insight and high-level information processing. Li and Lu [18] found Gamma waves were suitable for emotion reading with emotional still images as stimuli. There are many feature extraction techniques to characterize the EEG signals on each band. Fourier energy [19] is a choice, but the electric activity at each electrode is not a stationary process over time [20]. To solve this problem, researchers introduced window functions to analyze signals within a short span of time, in which the process could be thought to be stationary. Hadjidimitriou et al. employed spectrogram, Hilbert-Huang spectrum, and Zhao-Atlas-Marks transform to classify ratings of liking and familiarity [21]. Li et al. applied wavelet energy to compensate for the influence of the non-stationary of EEG series [22]. Wang et al. compared power spectrum, wavelet, and nonlinear dynamic feature to classify emotion states and indicated that power spectrum was superior to the other two features [23]. Sabzpoushan et al. used six types of features and then selected the top-ranked subset of them to characterize EEG signal for classification [24]. After a detailed comparison test, Lu et al. proved that differential entropy (DE) was the most accurate and stable feature for emotion recognition than traditional features, including power spectrum density (PSD), autoregressive parameters, fractal dimension, and sample entropy [25, 26].

Based on the aforementioned features, researchers have constructed numerous models for emotion reading, e.g., [27] used SVM to classify happiness, relaxation, sadness, and fear, and [28] introduced autoregression modeling to judge whether the subject was in the positive or negative emotion state on band. However, the models they used were shallow, and the low SNR was still the major frustration for classification. In order to reduce the influence of the uncertainness, EEG signal analysis often involves in feature selection procedures or hand-crafted signal-representing techniques, e.g., PCA and Fisher Projection. However, they are insufficient to excavate patterns in EEG signals as they ignore the original information such as channels. On the other hand, the cost of traditional feature selection methods increases quadratically with respect to the number of features considered [29]. Therefore, we need more powerful models to learn efficient representations for the EEG-based brain decoding.

Since the work of Hinton and Krizhevsky in 2012 [30], deep learning (DL) has dominated the machine learning field and showed advantage in some complex tasks [31, 32]. The most common structures include HCNN [30], SAE [33], and deep belief networks (DBN) [34]. In the traditional models, the feature configuration is fixed, and the classifiers spared no effort to fit the task according to the features. As a contrast, DL transforms the features from layer to layer dynamically in the hierarchical structure and automatically learn the optimized combination of them. In DL, the feature transformation and classifier training are no longer independent procedures, because it binds the classifier training and representational learning together, and thus yields the optimum matching of them. In practice, deep networks need a great deal of training cases as they have plenty of parameters. As EEG brings no harm to human body during long-term experiment, the data amount requirement is not a barrier.

Recently, some researchers have used DL to analyze EEG signals. Provost et al. proposed an automatic emotion recognition system on the basis of DBN in an audiovisual task and offered the performance comparison with SVM [35]. They find that the learned high-order nonlinear relationships provided by deep networks were more effective for emotion recognition. Israsena et al. calculated the PSD features of 32-channel EEG signals and applied PCA to extract the most important components of initial input features. They used SAE to learn the nonlinear representations and classified the level of valence and arousal. Their results indicate that SAE is better than SVM and naive Bayes classifiers [36]. In our work, we also use SAE as a comparison method with HCNN, and the PCA operation is removed to maintain the channel information. Graeser et al. proposed to discriminate different mental states according to EEG signals from six electrodes via a shallow HCNN [37]. They organized the EEG signals as 2-D maps where each column corresponded to one electrode, and each row was one sample recorded at an interval of time resolution. Although their electrode number was quite limited, they showed the advantage of DL in EEG-based recognition. The two dimensions were electrode and time; whereas our work, both dimensions refer to electrode. We will discuss it in detail in "Methods."

Some researchers have shown that the individual difference of EEG signals across subjects is remarkable [20, 38, 39]. From person to person, the EEG signal differs by a large scale. Rothkrantz et al. showed that the subject-specific EEG data contained enough information to recognize emotion state, but

there was still a lot of diversity among different people and circumstances [40]. Transfer learning attempts to solve this problem. In emotion recognition, transfer learning makes use of the knowledge of the source domains (previous subjects) to help improve the classification accuracy in the target domain (the present subject) [41]. A recent study [20] offers us a comprehensive survey on the transfer learning in BCIs. In the traditional models, the most popular algorithms are based on the common spatial pattern (CSP) [42–44], which aims at finding an invariant subspace to project all subjects' data to. In this way, the shared part of different subjects' data is preserved, and the difference is compensated. To our knowledge, seldom works have been conducted concerning the transfer methods in aBCIs. Lu et al. personalized the emotion classifier trained on the previous subjects to the present person with transductive parameter transfer (TPT) [39]. However, the transfer methods are combined with shallow models, e.g., SVM. In our work, we will discuss the most popular transfer method in DL, i.e., the fine-tune technique [45].

In this paper, we adopt DE to characterize EEG signals as it has been proven to be suitable for emotion decoding in previous studies [10, 25, 26]. In order to maintain the information contained in the positional relationship between the electrodes, we organize data from different channels as 2-D sparse maps to train HCNN classifiers on Delta, Theta, Alpha, Beta, and Gamma bands. We implement the recognition models with SVM and KNN and compare them with DL. We also compare the performance of two deep models (HCNN and SAE) in emotion decoding. We analyze the classification accuracy across frequency bands to locate the critical bands. To evaluate the transfer capability of HCNN parameters among different peoples with fine-tune technique, we define three training strategies. In the end, we visualize the evolution process of representations along the hierarchical structures in 2-D maps with PCA technique and discuss the information-processing mechanism inside HCNN.

The rest of this paper is organized as follows: "Methods" introduces the core methods we use. "Experiments" describes the processes in detail. "Results" is the evaluation of the experimentation. "Discussion" is the analysis of the results and "Conclusion" is the completion section of the study.

Methods

In this section, we first propose the general framework of HCNN-based emotion recognition system, and then introduce each module in sequence.

The General Framework for Emotion Recognition

The HCNN-based emotion recognition system is shown in Fig. 1. The typical aBCI is a closed-loop system [46], where feedbacks are involved. Emotion recognition is the key module in aBCI. For illustration, we draw feedback in the figure. The visual, together with auditory stimuli, is used to evoke the subject's emotion, and EEG signals are recorded on 62 electrodes. After some standard preprocessing procedures, we compute DE features on all channels at certain time interval and organize them as 2-D maps. After the sparse operation, we use HCNN to process the input maps and classify emotion states.

Preprocessing

The raw EEG data are downsampled to 200-Hz sampling rate. The signals are visually checked and the recordings seriously contaminated by EMG and EOG are removed manually. With the help of EOG recordings, the blink artifacts are located and removed. In order to filter the noise and remove the artifacts, the EEG data are processed with a bandpass filter between 0.3 and 50 Hz.

Short-Time Fourier Transform and Differential Entropy

Fourier transform (FT) is often used to analyze the frequency configuration of a time-domain signal, and it is widely used in EEG decomposition. However, the FT operation assumes that brain wave activities are stationary, which is a false hypothesis apparently. Therefore, the time series should be cut into small time segments, and within each segment, the brain electric activities are approximately considered as stationary. The idea is called the short-time Fourier transform (STFT). STFT decomposes a function of time (in our case, EEG signal) into the frequencies that make it up at fixed time intervals. The calculation formulation of STFT is as follows:

$$\mathrm{X}(\tau,\omega)=\int_{-\infty}^{\infty}x(t)\omega(t-\tau)e^{-j\omega t}dt, \tag{1}$$

where $x(t)$ is the original signal and $\omega(t)$ is the window function. The Hanning window (shown in (2), a discrete version) is a linear combination of modulated rectangular windows, and it usually emerges in applications that require low aliasing and less spectrum leakage.

$$\omega(\mathrm{n})=\frac{1}{2}\left(1-\cos\left(\frac{2\pi n}{N-1}\right)\right), \tag{2}$$

where n is the window length and N is the sampling number. DE [25] is referred to as the extension of the Shannon entropy

$$\mathrm{h}(\mathrm{x})=-\sum_{i=1}^{N}p(x_i)\log(p(x_i)). \tag{3}$$

DE is the continuous version of Shannon entropy and the original computing formulation could be written as the following:

$$\mathrm{h}(\mathrm{x})=-\int f(x)\log(f(x))dx. \tag{4}$$

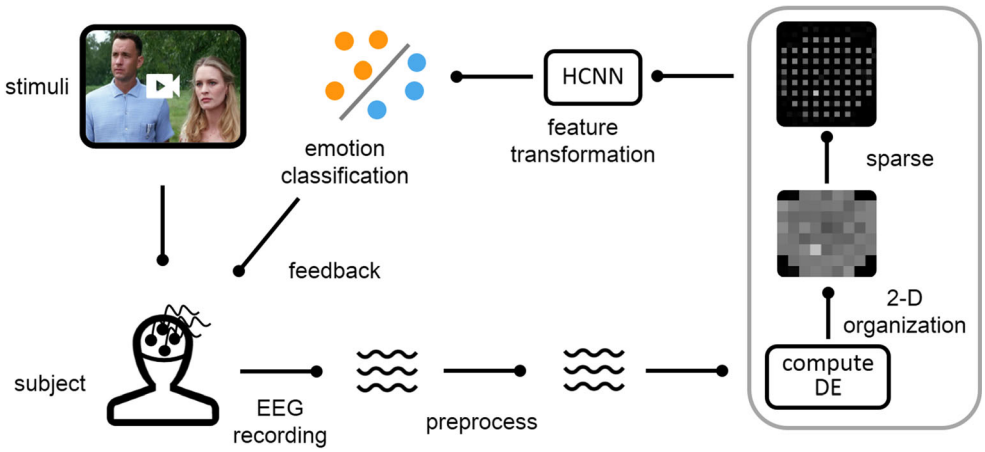

Fig. 1 The closed loop of aBCI system. Emotion recognition is the central task, and we implement it with HCNN

DE feature is a simple but efficient evaluation of the complexity of a discrete random variable (i.e., if we want to know a random variable thoroughly, how much information do we need). DE offers a good selection to characterize EEG time series. Previous studies have shown the advantage of DE over PSD [10]. According to [10], for a fixed length EEG signal, DE is equivalent to the logarithm of PSD in a certain frequency band. If a random variable obeys the Gaussian distribution $N(\mu, \sigma^2)$, the differential entropy in (4) can simply be calculated by the following formulation:

$$\begin{aligned} h(x) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2} \log\frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2} dx \\ &= \frac{1}{2}\log 2\pi e\sigma^2, \end{aligned} \tag{5}$$

where e is the Euler's constant and σ is the standard deviation of a time series. The complexity of the series is evaluated by its deviation.

Two-Dimensional Feature Organization

In order to maintain the information hidden in the electrode placement, we organize DE features extracted from 62 channels as 2-D maps at a time interval of 1 s. There are few sophisticated mapping techniques to map the electrodes to 2-D maps, e.g., azimuthal equidistant projection (AEP) known as polar projection borrowed from mapping applications [47]. In AEP, we are able to maintain the distance of each electrode to a certain central point in the 3-D sphere in the 2-D plane; however, the relative distances between the electrodes are discarded. In this paper, we organize the map in the simplest way, and the positional relationship is quite intuitive. The configuration of the DE map is illustrated in Fig. 2. We organize the DE features in such configuration to feed HCNNs.

The 2-D organization is inspired by the working principle of HCNN. In the HCNN structure, each convolutional kernel only has localized receptive field [30], so the kernels are able to capture the correlationship among adjacent electrodes, which might be of great value for the recognition task. In the final classification, the last layer's feature maps are all stretched to one-dimensional vectors and full-connected to the output neurons, so the global electrode placement is considered.

However, we should deal with two problems: first, for the 62-electrode placement, the map size is too small (8×9), so it is hard to train a HCNN with multiple layers. Second, in the 8×9 map, for each electrode, the feature value is closely surrounded by its adjacent electrodes' values, which means the features are too "compact" or "concentrated." This may result in large-scale information leakage in the immediately following convolution and pooling operations, especially in the pooling operations. Therefore, we introduce sparsity to generate sparse DE maps: all-zero rows and columns are added on alternate rows and columns. All-zero frames are also added on the four edges of maps to maintain patterns hidden in peripheral electrodes (see Fig. 1). In this way, the DE maps are more suitable for HCNN dispose, and the information leakage will be reduced. There are several techniques to add interpolate values at the zero-value positions, e.g., linear interpolation, Gauss interpolation, and Clough-Tocher scheme [48]. We do not apply the interpolation tricks to preserve sparsity and reduce unrelated noise. Therefore, after the sparse operation, the DE map is 20×20. The size is sufficient to train a small-scale HCNN.

Hierarchical Convolutional Neural Networks

HCNN is powerful in its representational ability; therefore, the features extracted in the hidden layers are often directly used as descriptors of input images [49, 50]. In this paper, the DE

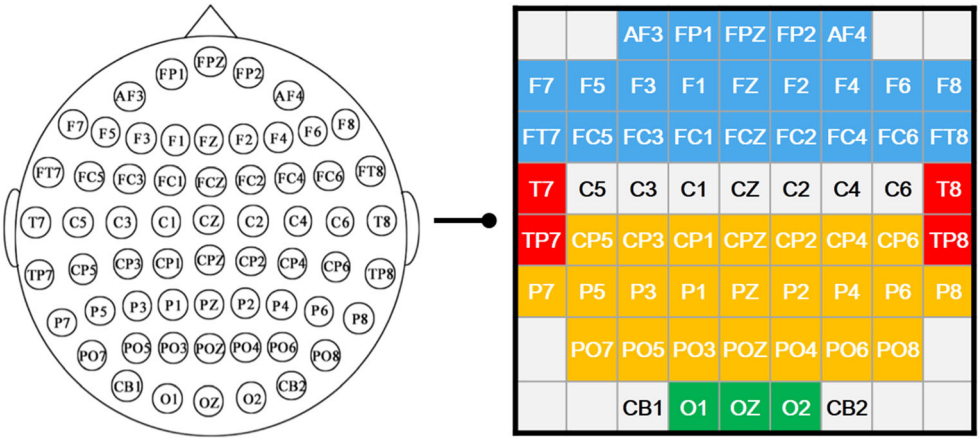

Fig. 2 The organization of DE map. Left: the 62-channel EEG placement in this experiment. Right: the placement is approximately expressed as a 2-D map, and each "pixel" corresponds to a certain electrode. Through this method, the spatial relationship of electrodes are taken into consideration. (*F* frontal, *T* temporal, *C* central, *P* parietal, *O* occipital, *Z* midline)

maps could be comprehended as the input images; hence, the emotion recognition task is transformed to a computer vision task. The HCNN plays two roles:

- As feature transformer. From layer to layer, the network disposes the input feature maps and project them to another space for better representation.
- As classifier. The transformed representations are used to classify the emotion state at the last full-connection layers.

The convolution operation is characterized by localized receptive field and the weight sharing [30]. The former endows the network with the ability to excavate patterns in the previous layer efficiently and locally via a relatively simple computation. The latter reduces the parameter amount remarkably. The pooling operations include average pooling (adopted in early HCNN structures while seldom used recently), max-pooling [51] (commonly used in recent DL research), and spatial pyramid pooling [52] (often emerge in multi-scale cases). Pooling is a type of dimensionality reduction method, which are also beneficial to control the parameter amount to reduce the computational burden. It also enhances the robustness and anti-noise capacity of the network.

The recognition accuracy of HCNNs with max-pooling unit in computer vision tasks is good, and scholars offer some physiological explanations: the one-win-all phenomenon is resemble to the neural functional mechanism in the human brain, and this characteristic might be valuable the machine learning tasks. We adopt max-pooling in this work.

In HCNN (see Fig. 3), from layer to layer, the representations become increasingly global and abstract, which is similar with the way the human visual cortex works (especially the primary visual cortex in the ventral pathway). The activation function and the pooling operation also endow the network with nonlinear feature transformation ability, which is vital in accomplishing complex visual goals.

Experiments

Subjects and Stimulus

We validate the effectiveness of our method on a large-scale EEG-based emotion recognition dataset called "SEED." It was originally contributed by Lu et al. [10, 26]. For more detailed information about it, or download it for scientific use, please log on to http://bcmi.sjtu.edu.cn/~seed/index.html.

The signals were recorded using an ESI NeuroScan System. There were 15 healthy subjects (8 females, mean 23.27, and SD 2.37) selected by the Eysenck Personality Questionnaire (EPQ). EPQ assesses the personality traits of a person [53]. All the subjects were with self-reported or corrected-to-normal vision and normal hearing. To exclude the influence of gender and reduce computational burden, we picked up 4 healthy male subjects. The subjects watched 15 emotion-evoking movie slices, which included 5 positive ones, 5 negative ones, and 5 neutral ones. The slices were cut from 6 movie sources. The movie sources are listed in Fig. 4. They were carefully selected to make sure that each slice corresponded to only one emotion state. The display order and durations of the slices are also shown in Fig. 4 (see the bottom). We added all the durations up, so the stimulus lasted 3394 s in total. If we take one sample per second, 3394 samples were thus obtained. During the movie-watching, 62-channel EEG signals were collected simultaneously.

Feature Extraction and Organization

Feature extraction and organization are significant. We assume that the signal is approximately a stationary process within each 1-s segment. In each segment, we perform 256-point STFT with non-overlapped Hanning window. DE is then calculated on wave band Delta, Theta, Alpha, Beta, and Gamma. For each of the five frequency bands, we organize the DE features as 2-D maps as described in "Methods." We also introduce sparsity by adding all-zero rows and columns to

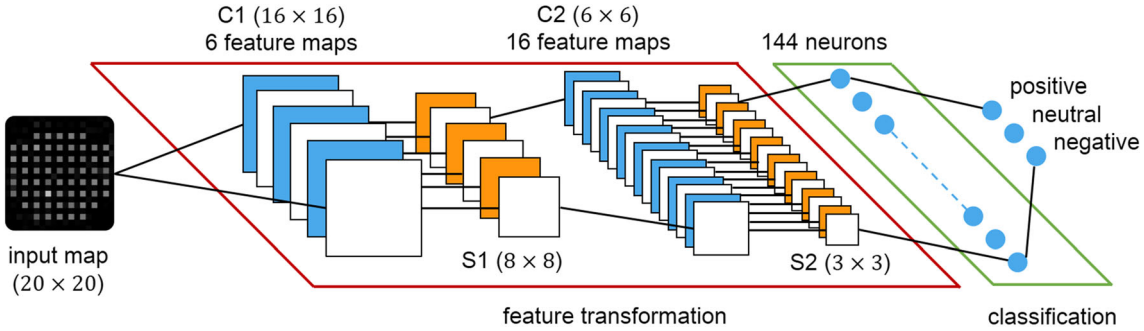

Fig. 3 The HCNN structure in the experiment. The first part: feature transformation. The second part: classification

the DE maps, which is beneficial to HCNN training. Finally, the sparse maps of each frequency band, together with the slice labels, are used as the teacher signals to train corresponding HCNN with supervised learning. At the end of this stage, we obtain five HCNNs. Each HCNN is special for one frequency band. Each band might contain emotion-related patterns, but the information amount is different. We will locate the most efficient band(s) for emotion recognition according to the classification accuracy.

Train HCNN

For each subject on each frequency band, we use 75% of the samples for training, and the remaining 25% for test (the test samples are not included in the training samples).

In practice, the HCNN configuration, including the network depth, the number of feature maps in each layer, and the kernel number for each layer should be carefully designed to adapt to input size and training goals. We implement a four-layer HCNN (shown in Fig. 3), which consists of two convolution layers and two max-pooling layers. The number of layers is limited because the input map size is small. If we concatenate extra layers in the downstream, the 2-D characteristic of the maps will vanish.

Movie slices sources	Label
Tangshan Earthquake	Negative
1942	Negative
Lost in Thailand	Positive
Flirting Scholar	Positive
Just Another Pandora's Box	Positive
World Heritage in China	Neutral

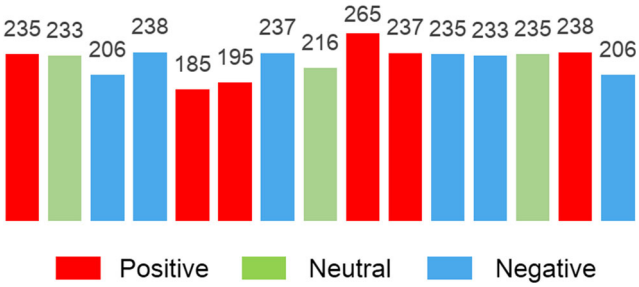

Fig. 4 The stimuli information. Up: the movie sources and labels. Down: the slice order (from left to right) and duration (in second)

In our experiment, the input DE map is 20 × 20. C1 is a convolution layer: the kernel size is 5, and the step is 1, then 6 16 × 16 maps are formed. S1 is a max-pooling layer, and the scale size is 2, so 6 8 × 8 maps are obtained. C2 is the second convolution layer: the kernel size is 3, and the step is 1. The input of C2 is the 6 maps in S1, and we appoint 16 filters. So the feature map number expands to 16. Therefore, 16 6 × 6 maps are formed in C2. S2 is another max-pooling layer, and scale size is also 2, so we get 16 3 × 3 maps in S2. All maps in S2 are then stretched and concatenated to generate a 144-D vector according to their spatial positions. The vector is then fully connected to 3 output nodes. Each node corresponds to one emotion state.

The C1, S1, C2, and S2 layers form the feature transformation module, where the sparse DE maps are processed gradually to generate better representations for the classification task. The full-connection layer performs categorization on the basis of the representations delivered up by the feature transformation module. The two modules work together.

The activate function is sigmoid. The initial values of the networks are randomly appointed. The learning rates are all set to 1 and the batch size is 50. The learning epoch is 600. All the experiments are done in MATLAB (R2014a) software.

Train SAE and Shallow Classifiers

To show the superiority of HCNN, we also implement the emotion classifiers with other models, i.e., the 1-D deep model SAE and the shallow models kernel-SVM and KNN. The performance comparison of HCNN with SAE illustrates the benefits of the topology-preserving 2-D feature organization. The comparison of deep models (HCNN and SAE) with shallow models (SVM and KNN) could show us the merits of deep representational learning.

SAE is the stacked architecture of several autoencoders. Each autoencoder is trained with unsupervised learning method aimed at reconstructing the input by the hidden layer where

the neuron is fewer than the input layer via BP algorithm. We call the mapping from the input layer to the hidden layer "encoding," and the mapping from the hidden layer to the input layer "decoding." The hidden layer is forced to learn the compressed representation of the input data. Therefore, as the cascaded chain of AEs, SAE learns deep and compressed representations of the input. The encoding-decoding framework is widely used for dimensionality reduction and deep network pre-training.

The configuration of SAE in this experiment is shown in Fig. 5. The input layer is composed of 400 neurons. The second layer h_1 is a hidden layer, where 200 neurons are assigned to reconstruct the input layer (the virtual reconstructed layers are colored in gray). The third layer h_2 is also a hidden layer, where 100 neurons are assigned to reconstruct h_1. The last layer is the output layer with three neurons, and each neuron corresponds to one kind of emotion state. During some preceding experiments, we find that sparsity is also beneficial for SAE training. Therefore, we do not take the 62-D feature as SAE input; but instead, we stretch the sparse map (20×20) as a 400-D vector to train the SAE. Another benefit of this operation is that the HCNN and SAE share the input dimensionality (400), so they compete on a same platform. The former works in 2-D plane, while the latter works in 1-D. The former has localized receptive field, but the latter does not.

For kernel-SVM, the penalty factor c and the parameter g are determined by grid search (the searching regions are all from 2^{-8} to 2^{8}) and three-fold cross validation. The kernel is radial basis function (RBF). SVM focus on the samples that are closed to decision boundaries, and thus gives the samples with the lowest confidence the highest confidence. The basic SVM deals with two-class categorization whereas we have there emotion states to recognize. Therefore, we train three binary classifiers between each of the two emotion states and use one-to-one voting strategy [54] for the final combination.

For KNN, we choose K value ranging from 1 to 500 to determine the predicted label of each test sample. The distance is defined as spatial Euclidean distance. The memory-based KNN algorithm is simple and intuitive, but the computation burden is relatively heavy.

Subject Transfer Via Fine-Tune Technique

To investigate the consistency of emotion-related electric responses across subjects recorded by EEG signal, we define three HCNN training strategies.

A. Train and test HCNN on data acquired from one same subject.
B. Train the HCNN on the data acquired from three subjects, and then test HCNN on a new subject.

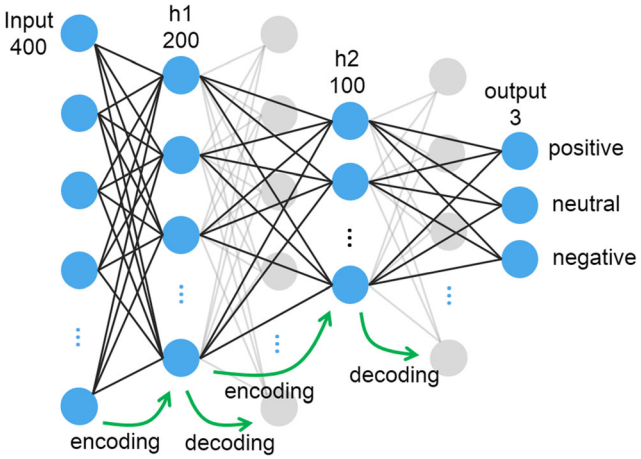

Fig. 5 The SAE architecture in our experiment. Each AE repeats the encoding and decoding operation

C. Pre-train the HCNN on the data of three subjects. Fine-tune it by part of the data of a new subject, and then test it on the new subject's remaining data.

Fine-tune is often used as an effective knowledge-transfer technique in deep neural networks. We introduce it into the emotion recognition networks and see to what extent the models (in the form of weights and bias of networks) could be shared by subjects. The comparison of A and B tells us the distribution discrepancy across subjects. The comparison of A and C might show us whether the initial values of HCNN parameters can enhance the recognition accuracy or boost the convergence. The comparison of B and C shows the effectiveness of few labeled data. The labeled samples are used to fine-tune parameters to adapt the network to new data distributions.

Feature Visualization along HCNN Layers

HCNN has nonlinear activation functions, and transforms the input DE map from layer to layer. Each layer outputs a new representation of the input map. Here, we offer a deep insight into the representations' evolution progress. This might help us understand why HCNN performs well. For simplicity, we take out the output of the two pooling layers (S1, S2) and apply linear dimensionality reduction (DR) technique to the high-dimensional features and plot them in 2-D maps. DR preserves the significant structure of the high-dimensional data in the low-dimensional data and could offer us an intuitive illustration.

We choose PCA to visualize the high-dimensional features. PCA is one type of linear DR method where the top two eigenvalues are preserved. In this work, the PCA operation aims at calculating the 2-D embedding of the high-dimensional feature space, and plot the data as points colored depending on their semantic categories in a particular

hierarchy. The linear separability on the 2-D plane could partly reflect the linear separability in the high-dimensional space.

Results

We first evaluate the complexity of the task with adjusted cosine similarity (ACS) criterion defined as

$$\mathrm{sim}(\mathrm{X},\mathrm{Y}) = \frac{\sum_{i=1}^{n}\left(X^i - A^i\right)\left(Y^i - A^i\right)}{\sqrt{\sum_{i=1}^{n}\left(X^i - A^i\right)}\sqrt{\sum_{i=1}^{n}\left(Y^i - A^i\right)}}, \quad (6)$$

where X and Y are two feature vectors with n dimensions. A is the mean of X and Y. Bigger value is equal to smaller angle, which means the two vectors are alike. We compute the mean ACS for each pair of the three emotion states on wave band Beta and Gamma. As Fig. 6 shows, the angles (in radian) are all close to zero. The features of different emotion states are alike. Classification is a hard task.

Table 1 shows the classification accuracies and time cost of the models. For KNN, we find the appropriate value of K is around 450 across all bands and subjects. No matter what kind of classifier is applied, Beta wave (best 86.2%) and Gamma wave (best 88.2%) show advantage over the others. Both HCNN and SAE are far better than SVM (best 74.4%) and KNN (best 79.3%).

On Beta, the mean accuracy of HCNN is 86.2%. It exceeds the accuracies of SAE, SVM, and KNN by 7.9%, 13% and 9.5%, respectively. On Gamma, the accuracy of HCNN is 88.2%. It beats SVM and KNN by a margin of 13.8% and 8.9%, respectively. However, the advantage of HCNN over SAE is not remarkable (2.8%).

The advantage of HCNN over SAE is relatively small on both Beta and Gamma, while the advantage over SVM and KNN is obvious especially on band Gamma. We will the discuss the reasons later.

Other than accuracy, computational efficiency (time cost) is another important portrait for algorithms. We run all the codes on an Intel (R) Core (TM) i7-4790 CPU with 3.6 GHz memory. The time efficiency ranking in the training process is KNN (no training), SVM, SAE, and HCNN. For SAE and HCNN, the price of higher accuracy is heavier computation burden during training. However, once the models are trained, we do not need to consider the training cost any more. In the test process, the memory-based KNN is the most time-consuming choice (~90 s), and the rest three algorithms make predictions less than 1 s. SVM, SAE, and HCNN are more suitable for real-time emotion reading than the memory-based KNN.

The performance of HCNNs trained under three strategies is summarized in Fig. 7. The horizontal axis specifies the training strategy. For strategy A, the training samples (75%) and test samples (25%) are on the same person. For B, all samples of the three previous subjects are used to train a HCNN, and 25% of the present subject's data are used for test. For strategy C, 75% of the present subject's samples are used for fine-tuning the network obtained in strategy B, and 25% samples are used for the final test. The mean accuracy of strategy A on the training data is as high as ~98% (not shown in the figures). Such results might make sense in retrieval applications. Our major concernment is a more difficult task, i.e., make prediction on new data. If HCNN trained on previous subjects is directly used on a new subject, some interesting phenomena draw our attention. Generally, the accuracies on Beta and Gamma show a downward cliff. On Beta, it is a surprising 50% decline. On Gamma, a 30% decline occurs. This might attribute to the distribution difference across subjects. However, Theta is a counterexample, where the classifier trained on strategy B is even better than the classifier trained on strategy A. This special case is not instructive since on Theta; the accuracy is close to chance (33%), which means Theta waves themselves do not contain much valuable information concerning emotion states.

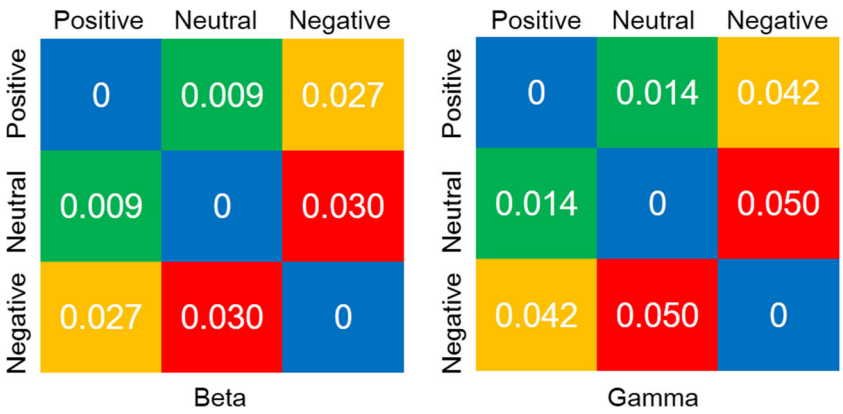

Fig. 6 The mean spatial angels for each pair of emotion states (on Beta band and Gamma band)

Table 1 The performance of SVM, KNN, SAE, and HCNN on five frequency bands

Classifier	Wave band					Time cost	
	Delta	Theta	Alpha	Beta	Gamma	Train	Test
SVM	.597(.062)	.501(.033)	.595(.155)	.732(.178)	.744(.097)	< 1 s	< 1 s
KNN	.559(.037)	.498(.049)	.587(.052)	.767(.133)	.793(0.82)	/	90 s
SAE	.491(.112)	.407(.090)	.671(.118)	.783(.132)	.854(.081)	350 s	< 1 s
HCNN	.369(.032)	.278(.106)	.390(.091)	.862(.066)	.882(.035)	3600 s	< 1 s

The accuracies are shown in "mean (SD)"

When fine-tune method (strategy C) is applied on the networks generated by B, the mean accuracy of Beta-classifier improves about 30%, reaching around 67%, but still lower than that of strategy A (86%). On band Gamma, we find the performance of strategy A wins C by 8%.

The above results have persuaded us that comparing with KNN, SVM, and SAE, HCNN is the best choice to categorize emotion states. Why HCNNs perform well? To shed light on this question, we visualize the feature-evolution progress of the hidden layers of HCNN on 2-D plane with linear-mapping PCA. The results are shown in Fig. 8.

For simplicity, we randomly choose one subject's data and train HCNNs (now we do not separate training samples and test samples, and all samples are used to train the networks). We provide the visualizations on band Beta and Gamma for S1 and S2. In Fig. 8, each map consists of more than 3000 dots, where each dot specifies a sample. The PCA visualization shows us some important clues about the feature transformation process along the layers. In the input space, the features of different categories are highly mixed and coupled together, and we cannot easily discriminate them by linear classifiers, which is also confirmed by ACS criterion. However, after a convolution computation and sigmoid mapping followed by max-pooling; in S1, the representations belonging to different categories are more linearly separable, as is shown in the middle row. Then the features go through another convolution layer and sigmoid mapping, as well as a max-pooling operation. In S2, the network yields more linearly separable representations (see the bottom row). Representation plays the key role in pattern recognition. Good representations make patterns easy to distinguish. HCNN generates appropriate representations. The representations make classifiers easy to train.

Figure 8 also shows two interesting phenomena in the 2-D projection of the representations:

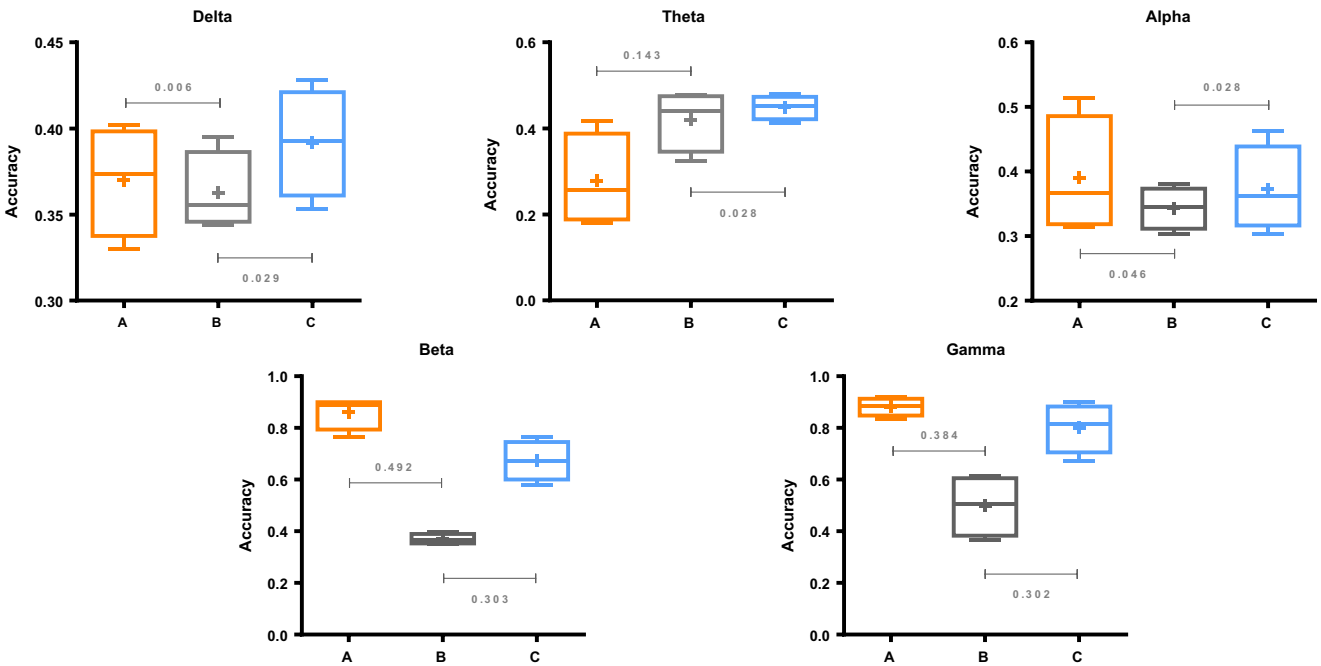

Fig. 7 The classification accuracies of HCNNs trained under strategy A, B, and C on five frequency, respectively. The accuracy difference is also offered

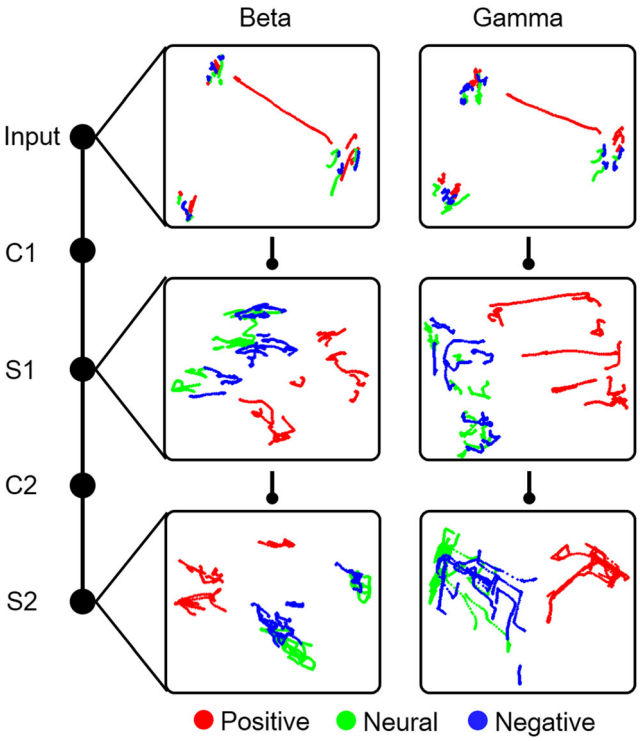

Fig. 8 The PCA feature visualizations of HCNN on Beta and Gamma bands. The top row is the original DE map, the middle row (S1) is the output of the first max-pooling layer, and the bottom (S2) is the output of the second max-pooling layer (best viewed in electronic version)

- Positive emotion is easy to separate. In each of the three rows, we find the positive emotion (red dots) is more easily to be separated from the neutral emotion (green dots) and negative emotion (blue dots), but the latter two are so similar to each other that it is hard to discriminate them. Therefore, the neural and negative emotion are similar in EEG recordings, while the positive emotion holds its own character. This is consistent with our daily experience. The positive emotion state usually means arousal and excitement, while the negative emotion states such as grief and sorrow are more closed to "no emotion" or "no arousal." The EEG signal reflects such traits.
- Positive and negative are more closed with each other in EEG. In each row, the red dots are more closed with blue dots, which means the EEG features of positive and negative emotion are more similar with each other. This is conflict with the intuitive imagination. We might think the "no emotion" should lie between positive emotion and negative emotion, but the reality is not so. The reason is simple: EEG is a reflection of "neural arousal" (the DE feature we use also measures the signal complexity). Positive and negative are both "evoked" emotion states where the arousal degree is naturally higher than "no emotion (neutral)."

Discussion

We conduct exhaustive experiments to investigate the use of HCNN in emotion classification. In this section, we discuss five noteworthy points.

First, deep models are better than the shallow models. When evaluating the similarity of EEG features of different emotion states, we find it hard to separate them. Emotion recognition itself is a difficult task. In this work, HCNN and SAE are deep models. In HCNN, the layer-wise convolution operations is efficient in feature synthesis and denoising. In SAE, the representational ability acquired through goal-driven BP algorithm in the encoding-decoding framework is strong. In our task, DL overbeats shallow models with obvious advantage. This phenomenon validates the capability of DL in complex data mining applications. The advantage of DL primarily lies in the powerful representational ability, which keeps growing as the models go deeper. The basic idea of DL is "deep" rather than "wide." It is quite similar to the visual information-processing mechanism in the human brain. For example, the visual information is not equally distributed across the entire cerebral cortex, but hierarchically distributed like a stream to form different functional-specific pathways. Each pathway is composed of several regions, and features are recomposed along the regions. DL works in the same way, and thus owns an insight into the data's internal structures.

Second, the topology-preserving 2-D organization is helpful to improve the performance. The accuracy of HCNN is better than SAE, which might be partly attributed to the way we organize the DE features. The 2-D maps contains extra (localized) spatial information for emotion recognition, and thus contains more information. Therefore, despite the training process has a demand on computational resource, HCNN is still a favorable tool to implement EEG-based emotion reading system. Here, we consider a confusing issue: the advantage of HCNN over SAE is not obvious because SAE also has the ability to excavate spatial information, especially when the network is deep. The 1-D SAE indeed learns the correlationship across the input dimensions, but the learned relationship miss some important information. We explain it in a more general case. Imagine a set of 2-D information is stretched into a 1-D vector, then the 2-D spatial information is hidden (more precisely, partly lost). If the space coordination information is not provided as well with the stretched 1-D data, we could not assert the 1-D and 2-D configurations are the same. The information leakage will occur for sure, and it is impossible to reconstruct the original 2-D data from the 1-D data alone. Therefore, the 2-D feature organization, as well as the HCNN structure (before the last few full-connected layers, HCNN transforms feature in 2-D geometrical architecture) are able to maintain the spatial

positional relationship among the adjacent electrodes, while the 1-D SAE structure could not. The good performance of SAE should be mainly attributed to its depth and learning method.

Third, Beta and Gamma bands are more suitable for emotion recognition. No matter how the features are organized and what kind of classifiers are applied, Beta waves and Gamma waves are always the best predictors of the emotion state. Gamma is slightly better than Beta in the average value, and the deviation is smaller (the performance is more stable). As a comparison, the accuracy obtained on Delta, Theta and Alpha are poor and unstable. The results demonstrate the significance of high-frequency brain waves in emotion recognition, and confirm previous findings that Beta waves and Gamma waves are highly-involved in the reasoning and thinking activities of the brain. Therefore, we should focus on these two bands in the investigation of emotion-processing mechanism and the implementation of aBCI systems.

Fourth, the parameter initial values learned on other people do not help the present HCNN to converge faster or better. In practice, this phenomenon tells us that the difference of EEG response patterns among different persons on the same task is considerable (even in the same nation and culture). The instance-based transfer method with SVM classifier is a favorable choice [39]. Transfer learning in emotion recognition is still a valuable research direction.

Fifth, in 2-D representation, the EEG recordings of the positive emotion is more unique, but the neutral and negative recordings are more alike. These findings is consistent with our experience that the positive emotion is linked with activity and zeal, whereas negative emotion is not far from "no emotion." Moreover, in EEG recordings, the positive emotion is close to negative emotion, and the neutral emotion does not sit between them. We should consider the emotion discrepancy from the view of "arousal," not from intuitive inferences.

In this paper, we train five HCNNs on five-frequency bands individually to see the performance difference and locate the best bands. Fusing the five HCNNs could result in better performance, which is a good direction for further research.

In the real world, the emotion stimuli could be movie, picture, music, language, and any other phenomena. Since the evoked emotion is the same, the emotion-related electric potentials are expected to hold consistence across modalities. The classifier might survive across different phenomena. This is another interesting research topic.

Conclusion

In this paper, we use various models to classify three emotion states. Benefitting from the powerful representational ability, deep models (HCNN and SAE) show absolute advantage over traditional shallow models. We find the high-frequency brain waves, i.e., Beta and Gamma waves, are highly involved in emotion processing. The visualizations of HCNN hidden layers show the feature evolution along the cascaded layers, where the representations become linearly separable. This phenomenon proves and explains the superiority of HCNN in emotion recognition.

Funding This work was supported by the National Natural Science Foundation of China (91520202, 81671651), CAS Scientific Equipment Development Project (YJKYYQ20170050) and Youth Innovation Promotion Association CAS. The authors would also like to thank Prof. Baoliang Lu forproviding the SEED dataset.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

1. Mühl C, Allison B, Nijholt A, Chanel G. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges[J]. Brain Comput Interfaces. 2014;1(2):66–84. https://doi.org/10.1080/2326263X.2014.912881.
2. Kothe CA, Scott M. Estimation of task workload from EEG data: new and current tools and perspectives. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2011.
3. Shi L-C, Bao-Liang L. EEG-based vigilance estimation using extreme learning machines. Neurocomputing. 2013;102:135–43. https://doi.org/10.1016/j.neucom.2012.02.041.
4. Sauvet F, Bougard C, Coroenne M, Lely L, van Beers P, Elbaz M, et al. In-flight automatic detection of vigilance states using a single EEG channel[J]. IEEE Trans Biomed Eng. 2014;61(12):2840–7. https://doi.org/10.1109/TBME.2014.2331189.
5. Ahern GL, Schwartz GE. Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis. Neuropsychologia. 1985;23(6):745–55. https://doi.org/10.1016/0028-3932(85)90081-8.
6. Gunes H, Piccardi M. Bi-modal emotion recognition from expressive face and body gestures. J Netw Comput Appl. 2007;30(4):1334–45. https://doi.org/10.1016/j.jnca.2006.09.007.
7. Busso C, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information. Proceedings of the 6th international conference on Multimodal interfaces. ACM, 2004.
8. Elfenbein HA, Ambady N. When familiarity breeds accuracy: cultural exposure and facial emotion recognition. J Pers Soc Psychol. 2003;85(2):276–90. https://doi.org/10.1037/0022-3514.85.2.276.
9. Russell JA. Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. Psychol Bull. 1994;115(1):102–41. https://doi.org/10.1037/0033-2909.115.1.102.
10. Zheng W-L, Bao-Liang L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural

networks. IEEE Trans Auton Ment Dev. 2015;7(3):162–75. https://doi.org/10.1109/TAMD.2015.2431497.
11. Hjorth B. EEG analysis based on time domain properties[J]. Electroencephalogr Clin Neurophysiol. 1970;29(3):306–10. https://doi.org/10.1016/0013-4694(70)90143-4.
12. Valdés P, Bosch J, Grave R, et al. Frequency domain models of the EEG[J]. Brain Topogr. 1992;4(4):309–19. https://doi.org/10.1007/BF01135568.
13. Davis CJ, Clinton JM, Jewett KA, et al. EEG delta wave power: an independent sleep phenotype or epiphenomenon[J]. J Clin Sleep Med. 2011;7(5)
14. Buzsáki G. Theta oscillations in the hippocampus[J]. Neuron. 2002;33(3):325–40. https://doi.org/10.1016/S0896-6273(02)00586-X.
15. Tran Y, Craig A, McIsaac P. Extraversion–introversion and 8–13 Hz waves in frontal cortical regions[J]. Personal Individ Differ. 2001;30(2):205–15. https://doi.org/10.1016/S0191-8869(00)00027-1.
16. Levin RB. Devices and methods for maintaining an alert state of consciousness through brain wave monitoring: U.S. Patent 6,167,298[P]. 2000–12-26.
17. Başar-Eroglu C, Strüber D, Schürmann M, Stadler M, Başar E. Gamma-band responses in the brain: a short review of psychophysiological correlates and functional significance[J]. Int J Psychophysiol. 1996;24(1):101–12. https://doi.org/10.1016/S0167-8760(96)00051-7.
18. Li Mu, Bao-Liang Lu. Emotion classification based on gamma-band EEG. 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2009.
19. Ferree TC, Hwa RC. Power-law scaling in human EEG: relation to Fourier power spectrum[J]. Neurocomputing. 2003;52:755–61.
20. Jayaram V, Alamgir M, Altun Y, Scholkopf B, Grosse-Wentrup M. Transfer learning in brain-computer interfaces[J]. IEEE Comput Intell Mag. 2016;11(1):20–31. https://doi.org/10.1109/MCI.2015.2501545.
21. Hadjidimitriou SK, Hadjileontiadis LJ. EEG-based classification of music appraisal responses using time-frequency analysis and familiarity ratings[J]. IEEE Trans Affect Comput. 2013;4(2):161–72. https://doi.org/10.1109/T-AFFC.2013.6.
22. Li Y, Luo M-L, Li K. A multiwavelet-based time-varying model identification approach for time–frequency analysis of EEG signals. Neurocomputing. 2016;193:106–14. https://doi.org/10.1016/j.neucom.2016.01.062.
23. Wang XW, Nie D, Lu BL. Emotional state classification from EEG data using machine learning approach[J]. Neurocomputing. 2014;129:94–106. https://doi.org/10.1016/j.neucom.2013.06.046.
24. Bajoulvand A, Marandi RZ, Daliri MR, et al. Analysis of folk music preference of people from different ethnic groups using kernel-based methods on EEG signals[J]. Appl Math Comput. 2017;307:62–70.
25. Duan, R-N, Zhu J-Y, Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on. IEEE, 2013.
26. Shi LC, Jiao YY, Lu BL. Differential entropy feature for EEG-based vigilance estimation[C]. Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. IEEE, 2013: 6627–30.
27. Lokannavar S, et al. Emotion recognition using EEG signals. Emotion. 4(5):2015.
28. Kahn B. Electroencephalogram (EEG) signal processing, wave identification, and emotion recognition. Diss. California State University, Northridge, 2015.
29. Dash M, Liu H. Feature selection for classification. Intelligent Data Analysis. 1997;1(3):131–56. https://doi.org/10.1016/S1088-467X(97)00008-5.
30. Krizhevsky A, Ilya Sutskever, GE Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012.
31. Szegedy C, et al. Going deeper with convolutions. Proc IEEE Conf Comput Vis Pattern Recognit. 2015;
32. Deselaers T, et al. A deep learning approach to machine transliteration. Proceedings of the Fourth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2009.
33. Vincent P, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. 2010;11:3371–408.
34. Hinton GE. Deep belief networks. Scholarpedia. 2009;4(5):5947. https://doi.org/10.4249/scholarpedia.5947.
35. Kim, Yelin, Honglak Lee, Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. 2013 I.E. International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.
36. 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;2014:1–10. https://doi.org/10.1155/2014/627892.
37. Cecotti H, Graeser A. Convolutional neural network with embedded fourier transform for EEG classification. Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008.
38. Samek W, Meinecke FC, Müller KR. Transferring subspaces between subjects in brain–computer interfacing[J]. IEEE Trans Biomed Eng. 2013;60(8):2289–98. https://doi.org/10.1109/TBME.2013.2253608.
39. Zheng W-L, Lu B-L. Personalizing EEG-based affective models with transfer learning, to appear in Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16), New York.
40. Horlings R, Datcu D, Rothkrantz LJM. Emotion recognition using brain activity. Proceedings of the 9th international conference on computer systems and technologies and workshop for PhD students in computing. ACM, 2008.
41. Pan SJ, Yang Q. A survey on transfer learning[J]. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59. https://doi.org/10.1109/TKDE.2009.191.
42. Fazli S, Popescu F, Danóczy M, Blankertz B, Müller KR, Grozea C. Subject-independent mental state classification in single trials[J]. Neural Netw. 2009;22(9):1305–12. https://doi.org/10.1016/j.neunet.2009.06.003.
43. Kang H, Nam Y, Choi S. Composite common spatial pattern for subject-to-subject transfer[J]. IEEE Signal Processing Letters. 2009;16(8):683–6. https://doi.org/10.1109/LSP.2009.2022557.
44. Lotte F, Guan C. Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms[J]. IEEE Trans Biomed Eng. 2011;58(2):355–62. https://doi.org/10.1109/TBME.2010.2082539.
45. Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks?[C]//Advances in neural information processing systems. 2014: 3320–3328.
46. Wu D, Courtney CG, Lance BJ, Narayanan SS, Dawson ME, Oie KS, et al. Optimal arousal identification and classification for affective computing using physiological signals: virtual reality stroop

task[J]. IEEE Trans Affect Comput. 2010;1(2):109–18. https://doi.org/10.1109/T-AFFC.2010.12.

47. Snyder, JP. Map projections–A working manual. Vol. 1395. US Government Printing Office, 1987.
48. Alfeld P. A trivariate Clough—Tocher scheme for tetrahedral data. Comput Aided Geometric Des. 1984;1(2):169–81. https://doi.org/10.1016/0167-8396(84)90029-3.
49. Sharif Razavian, Ali, et al. CNN features off-the-shelf: an astounding baseline for recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014.
50. Donahue, Jeff, et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. ICML. 2014.
51. Nagi J, Ducatelle F, Di Caro G A, et al. Max-pooling convolutional neural networks for vision-based hand gesture recognition[C]// Signal and Image Processing Applications (ICSIPA), 2011 I.E. International Conference on. IEEE, 2011: 342–7.
52. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European conference on computer vision. Springer International Publishing, 2014: 346–361.
53. Eysenck SBG, Eysenck HJ, Barrett P. A revised version of the psychoticism scale[J]. Personal Individ Differ. 1985;6(1):21–9. https://doi.org/10.1016/0191-8869(85)90026-1.
54. Weston J, Watkins C. Multi-class support vector machines[R]. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May, 1998.