Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring

Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring

Xiang Li

Tianjin Key Laboratory of Cognitive Computing and Application,
School of Computer Science and Technology,
Tianjin University,
Tianjin 300350, China
Email: xlee@tju.edu.cn

Dawei Song*

Tianjin Key Laboratory of Cognitive Computing and Application,
School of Computer Science and Technology,
Tianjin University,
Tianjin 300350, China
and
School of Computing and Communications,
The Open University,
Milton Keynes MK76AA, UK
Email: dwsong@tju.edu.cn

Peng Zhang* and Yuexian Hou

Tianjin Key Laboratory of Cognitive Computing and Application,
School of Computer Science and Technology,
Tianjin University,
Tianjin 300350, China
Email: pzhang@tju.edu.cn
Email: yxhou@tju.edu.cn
*Corresponding authors

Bin Hu

The Ubiquitous Awareness and Intelligent Solutions Lab,
School of Information Science and Engineering,
Lanzhou University,
Lanzhou 730000, China
Email: bh@lzu.edu.cn

Abstract: How to fuse multi-channel neurophysiological signals for emotion recognition is emerging as a hot research topic in community of Computational Psychophysiology. Nevertheless, prior feature engineering based approaches require extracting various domain knowledge related features at a high time cost. Moreover, traditional fusion method cannot fully utilise correlation

information between different channels and frequency components. In this paper, we design a hybrid deep learning model, in which the 'Convolutional Neural Network (CNN)' is utilised for extracting task-related features, as well as mining inter-channel and inter-frequency correlation, besides, the 'Recurrent Neural Network (RNN)' is concatenated for integrating contextual information from the frame cube sequence. Experiments are carried out in a trial-level emotion recognition task, on the DEAP benchmarking dataset. Experimental results demonstrate that the proposed framework outperforms the classical methods, with regard to both of the emotional dimensions of Valence and Arousal.

Keywords: affective computing; CNN; time series data analysis; EEG; emotion recognition; LSTM; multi-channel data fusion; multi-modal data fusion; physiological signal; RNN.

Reference to this paper should be made as follows: Li, X., Song, D., Zhang, P., Hou, Y. and Hu, B. (2017) 'Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring', *Int. J. Data Mining and Bioinformatics*, Vol. 18, No. 1, pp.1–27.

Biographical notes: Xiang Li is currently pursuing the PhD degree in Computer Application Technology at Tianjin University since 2014. He received the BS and MS degrees in Network Engineering and Computer Application Technology from Shandong University of Science and Technology, China, in 2011 and 2014, respectively. His research interests lie at the intersection between computer science and brain informatics, including neurophysiological signal processing and modelling, affective computing, data mining and deep learning.

Dawei Song joined Tianjin University as a Professor under the Tianjin 1000-talent scheme in 2012. Prior to this appointment, he worked as a Chair in Computing since 2008 at the Robert Gordon University, UK, where he remained as an Honorary Professor since June 2012. He has also worked as a Senior Lecturer and Research Director (since 2007) at the Knowledge Media Institute of The Open University, UK, during 9/2005–10/2008; and as a Research Scientist (since 9/2000) and Senior Research Scientist (since 5/2002) at the Cooperative Research Centre in Enterprise Distributed Systems Technology, Australia. His research interests include information retrieval, text mining and intelligent system.

Peng Zhang is currently an Associate Professor at School of Computer Science and Technology, Tianjin University. He severed as the PC member for a number of IR major conferences, such as SIGIR, CIKM, ECIR and ICTIR, etc., and the reviewer for refereed journal articles. His current research interests include information retrieval, cognitive computing and machine learning.

Yuexian Hou is currently a Professor at School of Computer Science and Technology, Tianjin University. He is also the Associate Director of the Tianjin Engineering Research Center of Big Data on Public Security, and the Director of the Institute of Computational Intelligence and Internet Application, Tianjin University. He served as the PC member of a series of academic conferences, e.g., ECIR, ICIC and WCICA. His main research interests include machine learning, information retrieval and natural language processing.

Bin Hu is Professor, Dean of School of Information Science and Engineering, Lanzhou University, the leader of Intelligent Contextual Computing Group in Pervasive Computing Centre, Reader, Birmingham City University, visiting

Professor in Beijing University of Posts and Telecommunications, China and at ETH Switzerland. His research interests include pervasive computing, psycho-physiological computing, cooperative work and semantic web.

This paper is a revised and expanded version of a paper entitled 'Emotion Recognition from Multi-Channel EEG Data through Convolutional Recurrent Neural Network' presented at the 'IEEE International Conference on Bioinformatics and Biomedicine (BIBM)', Shenzhen, China, 15–18 December 2016.

1 Introduction

Emotion, which is sometimes also referred to as affect or mood, is the internal experience (e.g., joy, grief, scare, anger, sympathy, disappointment, etc.) caused by various external events. It plays important roles in various aspects for human beings, such as interpersonal communication, planning, creativity, reasoning, behaviour, etc. Increasingly, emotion is being regarded as an important part of intelligence, as the process of judgment, reasoning and decision making cannot avoid the influence of inner emotions (Blanchette and Richards, 1994). Therefore, lots of researchers in Artificial Intelligence (AI) believe that the machine cannot acquire a true intelligence without cognition of emotion (Minsky, 1988). In this context, a new research branch in AI called Affective Computing (AC) emerged. The goal of AC is to empower computer systems with the ability to recognise, comprehend, express and respond appropriately to human's emotions, and ultimately providing us with enhanced experience in various scenarios of Human Computer Interaction (HCI) (Picard, 1999), such as computer games (Liu et al., 2009), information retrieval (Lopatovska and Arapakis, 2011), safety driving assist system (Hernandez et al., 2014), e-learning (Shen et al., 2009), etc.

Besides the practical applications in HCI, emotion recognition and monitoring are also promising in the field of auxiliary diagnosis of various mood disorders. It is estimated by the World Health Organization (WHO) that major depression will be the second leading cause of disability in the world by 2020, trailing only ischemic heart disease (World Health Organization, 2001). There is an urgent need to develop effective techniques to assist the psychiatrists in diagnosing and precaution of emotional disorders, e.g., depression, post-traumatic stress disorder, anxiety, etc. Traditionally, the patients' mental status is clinically assessed by psychiatrists based on the DSM checklist (American Psychiatric Association, 2013) or through patients' self-reporting information. Nevertheless, the diagnosis accuracy may be affected by the proficiency of the psychiatrists and the cooperation of the patients. Therefore, a reliable computer aided emotion monitoring method will contribute largely to mental illness prevention and diagnosis.

Broadly speaking, emotion recognition is a pattern recognition task based on monitoring and analysing human's various physical manifestations. In general, explicit physical activities (e.g., facial expressions, body gestures and voice) are subconsciously manifested. They are the main ways to communicate our emotional information and personal feelings to other people. However, in some situations we may consciously conceal, pretend, even exaggerate those emotional manifestations. In this regard, those explicit manifestations based recognition approaches maybe not stable and robust in

performance. According to various psychophysiological studies, even though there are still debates on whether the emotion fluctuations prior to neurophysiological changes or oppositely neurophysiological changes result in various emotions, there definitely exists emotion-specific patterns that we cannot consciously control (Ekman and Davidson, 1994; Krumhansl, 1997; Louis and Laurel, 2001). More specifically, the generation of emotion reflects a synergy of the Central Nervous System (CNS) and the Automatic Nervous System (ANS). It suggests that we can explore emotion computational method based on neurophysiological measurements directly or indirectly related to CNS (e.g., EEG, fMRI, etc.) or ANS (e.g., blood volume, ECG, galvanic skin response, respiration, temperature, etc.).

In computational psychophysiology, limited by the development of acquisition apparatuses and computational methods, early researches mainly focused on single-channel neurophysiological signal based emotion recognition. With the recent advances in sensor technology and machine learning methods, synchronised monitoring, recording and analysing of multi-channel neurophysiological signal is becoming possible. Nevertheless, there still exist a number of challenges in multi-channel neurophysiological signal based emotion recognition and monitoring, summarised as follows:

- First, numerous efforts have been devoted to finding and designing the various emotion-related features from the weak and noisy multi-channel signals, such as differential entropy (Duan et al., 2013), asymmetric spatial pattern (Huang et al., 2012), nonlinear dynamic characteristics (Stam, 2006). However, the design of those features needs more in-depth study from the perspective of emotion related psychophysiological research. The computation of those features sometimes can be time consuming, especially when extracting features for a relatively long time series based on theory of chaos and non-linear dynamics, e.g., Laypunov exponent, and the effectiveness of those features are largely affected by the parameter settings, e.g., the length of the computing window, the number of embedding dimensions, the length of delay time, etc. (Kim et al., 1999). Meanwhile, various feature selection and reduction methods are needed and studied for finding critical emotion-related neurophysiological information and better recognition performance (Li et al., 2016; Zheng et al., 2016).
- Second, capturing the correlation between multi-channel neurophysiological signal is crucial, but typically done through feature-level fusion based approaches (Chen et al., 2015) or decision fusion based strategies (Hussain et al., 2011). However, the processes of feature extraction and correlation modeling in these methods are handled separately. Recent Deep Learning (DL) based shared representation learning approaches (Jirayucharoensak et al., 2014; Li et al., 2015), although promising, still largely rely on hand-engineered features, which may under-utilise the ability of DL, in which the task-related features and shared representation can be learned automatically.
- Third, traditional machine learning based methods are not good at modelling the transition and evolution of emotional states, which is important for psychiatrists to monitor, review and assess a patient's past condition. In addition, the traditional off-line machine learning methods are not suitable for incremental learning scenarios, where multi-channel neurophysiological signals may be continuously acquired online rather than be provided in advance.

- Fourth, most existing works have concentrated on segment-level emotion recognition tasks, where the emotion prediction is conducted for signal segments of 1 second long or a little longer. It cannot meet the need of long-term emotion monitoring, as the acquired signals may last for hours or days.

To address the issues mentioned above, we propose a pre-processing method that encapsulates the multi-channel neurophysiological signals into grid-like frame cubes. Each frame cube represents the wavelet spectral energy information of the multi-channel signals within a specific time window. Further, we propose a hybrid deep learning structure that integrates the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for processing the acquired frame cube sequences and conducting emotion recognition tasks in one single framework. Specifically, the CNN is used for learning task-related features and mining inter correlation of multiple-channel neurophysiological signals from the frame cube, through designed convolutional filters. The characteristics of deep learning in automatic feature extraction and feature selection reduce the difficulties in computing domain-specific features as well as help us to bypass the traditional feature selection phase. We can also determine the critical emotion-related neurophysiological variables and components through analysing the trained convolutional filters. The RNN is used for modelling the evolution, transition and long term dependencies of the signals for final emotion prediction. The adoption of RNN is based on two facts: (1) the emotional experience is a reaction to external events and evolves continuously with respect to the change of stimuli, and (2) the neurophysiological signals contain rich contextual and semantic information that is suitable for the RNN to model. In the experiment, we validated the effectiveness of our method on the multi-channel EEG data in DEAP dataset, which is a widely used benchmark for emotion recognition, and also shown that our method has a potential for real-time prediction and monitoring.

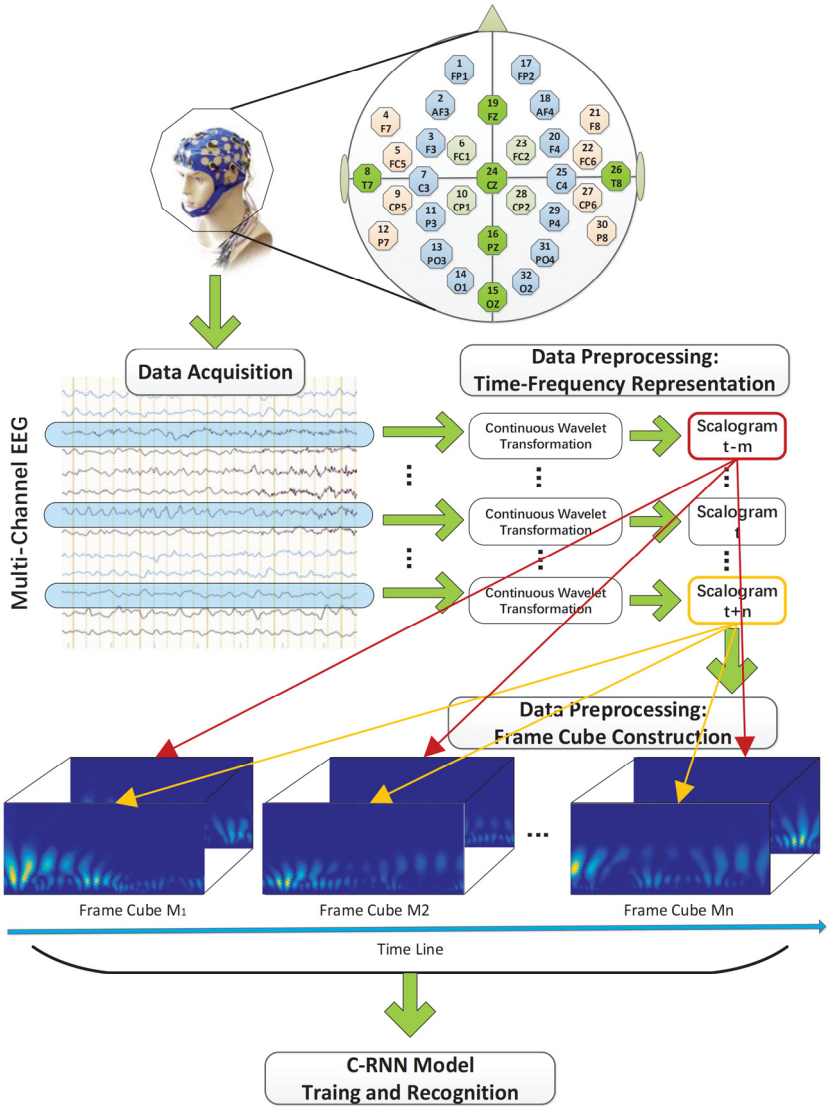
The rest of this paper is organised as follows. A detailed description of the proposed pre-processing method and the hybrid DL structure is presented in Section 2. Section 3 describes the experimental settings and reports the experimental results, as well as analyses the critical variables in emotion recognition, finally, we conclude this work in Section 4.

2 The proposed deep fusion framework

2.1 Overview of the framework

Our proposed methodology addresses two research problems. The first one is how to pre-process and represent the multi-channel signals before adopting some modelling methods. The second one is how to model and recognise emotions based on the pre-processed data. Correspondingly, we propose a 3D frame cube based representation method and a hybrid deep learning framework based on CNN and RNN, respectively. The brief process of our framework in dealing with multi-channel neurophysiological signal based emotion recognition is illustrated in Figure 1.

Figure 1 The overview of our framework designed for multi-channel neurophysiological data based emotion recognition (see online version for colours)



2.2 *Data pre-processing and frame cube construction method*

In this work, we construct three-dimensional data structure which is called 'frame cube' by us, denoted as M_n, each represents the wavelet energy distribution in dimensions of 'channel', 'scale' (a terminology in signal processing, each scale represents a specific coarseness of the signal) and the current time window. The advantage of the frame cube structure is that it encapsulates and integrates information of all channels' signal into a direct-viewing form. Therefore, these multi-channel signals can be further processed as a whole, and the inner relationship of the multiple channels can be mined, especially

suitable for the multi-channel EEG signal based data mining tasks, where each channel's signal is the mixture of electrical activity arising from various cerebral regions distributed in the brain. Furthermore, the sequence of frame cubes < M_1, M_2,..., M_n > reflects the dynamic activity changing in different cortical areas during an emotional experience.

Before constructing the frame cubes, we firstly need to conduct 'Continuous Wavelet Transform (CWT)' for each-channel signal, and we then need further transform the output from CWT into scalograms, as detailed below.

2.2.1 Wavelet based sparse representation

The representation and approximation of a signal is the key issue in signal processing and related pattern recognition tasks. The 'Sparse Representation (SR)' plays important roles in fields including signal processing, machine learning, computer vision, etc. It helps learn a compact structure and get high-level implicit semantic information from raw signals (Rubinstein et al., 2010). 'Wavelet Transform (WT)' is typically regarded as a kind of SR. It is excellent in denoting local transitory characteristics in both frequency and time domain. Therefore, it is quite suitable to be applied to process non-stationary neurophysiological signals, such as EEG (Subasi, 2005). Compared with the wavelet transform, traditional 'Windowed Fourier Transform' based time-frequency analysis (e.g., the 'Short Time Fourier Transform (STFT)') is only suitable for processing stationary signal and it is incapable of getting a high joint frequency-time resolution according to the 'Heisenberg's Uncertainty Principle'. Therefore, even though STFT is a good solution in some cases, in this work we choose to adopt wavelet transform.

Compared with traditional SR methods, such as 'Sparse Coding'. The dictionary of wavelet analysis is not acquired through learning, but is predetermined by a mother wavelet ψ. After scaling s and translation u of the mother wavelet a group of wavelet basis functions $\psi_{s,u}$ can be acquired, as Formula 1.

$$\psi_{s,u}(t) := \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right), u \in \mathbb{R},\ s > 0. \tag{1}$$

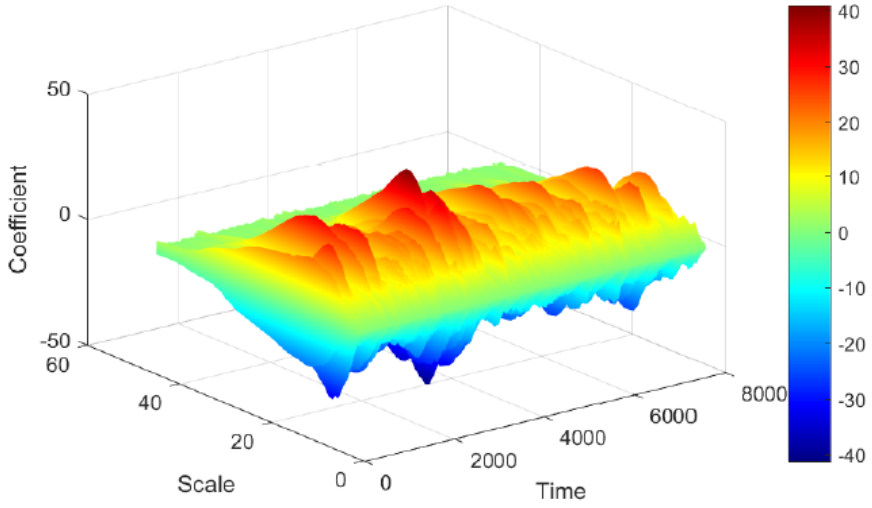
The raw channel signal is decomposed according to these basis functions. Therefore, the effect of representation is largely affected by what kind of mother wavelet is selected. You have to deliberately choose from various kinds of wavelet families, such as Haar wavelet, Daubechies wavelet, Symlet wavelet, Coif Wavelet, Bior Wavelet, etc. (Gandhi et al., 2011). In this work we simply choose the Db-4 wavelet to do CWT for each channel signal, which is formulated as Formula 2, where $f(t)$ is the original EEG signal.

$$W f(s,u) := \int_{-\infty}^{+\infty} f(t)\overline{\psi}_{s,u}(t)\,dt. \tag{2}$$

After the CWT, each one-dimensional channel signal is transformed into a wavelet coefficients based time-scale representation, as shown in Figure 2. The notion of scale is introduced as an alternative of frequency. Each scale corresponds to a scaled version of the mother wavelet, where the low scale is generally corresponding to high-frequency component of the signal and the high scale is corresponding to low-frequency component. The conversion relationship between specific frequency and scale is determined based on the sampling period and central frequency of the wavelet. The number of scales we specify is determined according to the properties of the raw signal,

such as the sampling rate and the cut-off frequency. For example, if the sampling rate is 128Hz, then we can obtain at most 64 frequency components from the raw signal (according to the 'Nyquist's Sampling Theorem'). Each element in the time-scale representation is the calculated wavelet coefficient corresponding to a specific scale of mother wavelet. The wavelet coefficients can also be used to reconstruct the original signal, so the wavelet coefficients here can be regarded as the signal's alternative representation.

Figure 2 The time-scale representation of the wavelet coefficients obtained after CWT for a channel signal. The larger the absolute value of the wavelet coefficient, the greater the proportion of the corresponding component in this channel signal (see online version for colours)

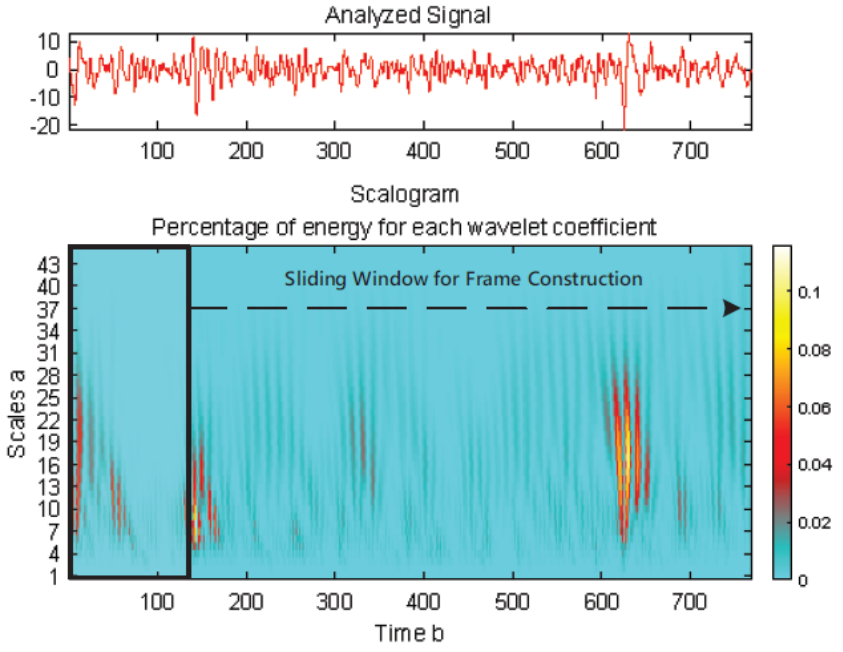


In this work, after the CWT we further transform each channel signal's wavelet coefficients based time-scale representation into an energy based time-scale representation, namely 'scalogram' (Bolos and Benitez, 2014), which can be obtained through Formula 3:

$$S(s) := \left(\int_{-\infty}^{+\infty} |Wf(s,u)|^2 \, du \right)^{\frac{1}{2}}. \tag{3}$$

As shown in Figure 3, it represents the distribution of the spectral energy in a signal, the hotter the pixel's colour the more concentration of energy in specific frequency range. The advantage of scalogram includes two aspects: First, the scalogram reflects detailed change of spectral in both time and scale, and the spectral energy oscillations have been recognised as an indicator of various cognitive processes (Ward, 2003). Second, each element of the scalogram is the percentage of energy that the corresponding signal component carries, and the sum of all the elements is equal to 1. Therefore, the numerical range is naturally suitable for processing by 'Artificial Neural Networks (ANN)'.

Figure 3 We further transform one channel signal into a scalogram based on the obtained wavelet coefficients, and the multi-channel scalograms are encapsulated into multi-time-step frame cubes through sliding window (see online version for colours)



We further construct frame cube sequence from the multi-channel scalograms. The reason for dividing the scalogram into multiple frames is the time window will be large if we represent the trial as only one single frame, especially when the signal's sampling rate is high. The processing of such a large-size input through CNN will lead to high computational cost. The cost in processing such a single large frame cube can be decreased by distributing the computing into smaller subparts, namely the frame cube sequence. The construction method for frame cube sequence is detailed in the following subsection.

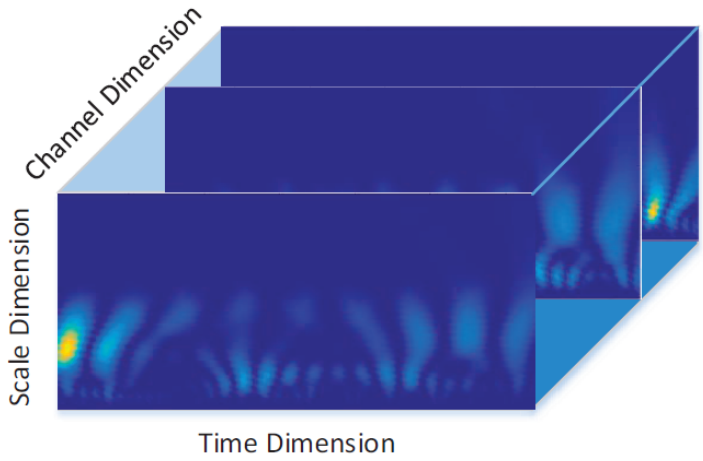
2.2.2 Frame cube construction

The frame cubes are constructed after each channel signal's scalogram has been obtained. Each frame cube is a cube-like $C \times S \times L$ structure, as in Figure 4, which represents the spectral energy distribution in the C channels and the S selected scales within a L-length time window. The procedure of constructing a frame cube can be summarised into three main operations:

- Firstly, we should determine the length L of the time window that a frame cube represents. For example, if we set the time window as 1 second long with no overlap between adjacent windows, then we can get 60 frame cubes for a 60 second long trial.
- Secondly, we extract frames by sliding the time window from start point simultaneously on each of those scalograms, each with a size of $S \times L$.

- Thirdly, we stack the extracted frame of each channel signal in current time step t to construct a frame cube. Then we can get a frame cube with a size of $C \times S \times L$, which represents the energy distribution within current time step t.
- Finally, we slide to the next time step $t + 1$ through sliding right with a length L, and repeat the operations from 1 to 3 until all of the frame cubes of one trial have been constructed.

Figure 4 The illustration of the proposed frame cube structure for information representation and mining (see online version for colours)



2.2.3 Scale selection

In order to reduce the computing burden, sometimes we also can adopt 'Energy to Shannon Entropy Ratio (EER)' to select some of the most representative scales. The optimal scales are selected when its spectral energy is high meanwhile its Shannon entropy is low. The criteria is presented as Formula 4:

$$r(s) = \frac{Energy(s)}{Entrophy_{sh.}(s)} \tag{4}$$

The energy of the scale 's' can be calculated through the sum of the energy that the 'n' wavelet coefficient of this scale carries, as Formula 5:

$$Energy(s) = \sum_{i=1}^{n} |C_i(s)|^2 \tag{5}$$

The Shannon entropy describe the uncertainty of the energy distribution in scale 's'. The lower the entropy the more information the specific scale contains, as Formula 6:

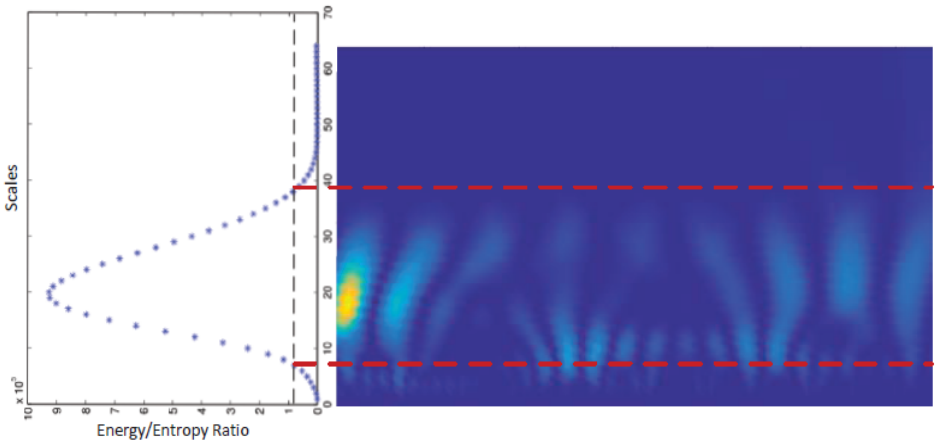
$$Eneropy(s) = -\sum_{i=1}^{n} P_i \log P_i \tag{6}$$

where the P_i is the probability distribution of the energy of the coefficient C_i in scale 's' and $\sum_{i=1}^{n} P_i = 1$, as Formula 7:

$$P_i = \frac{|C_i(s)|^2}{Energy(s)} \tag{7}$$

The EER for all channel signals can be calculated according to the method mentioned above. After averaging the results we got the average EER for each scale, as shown in Figure 5. For example, in this case we can select components in scale from 7 to 38 whose ratio is relatively high to construct the frame cubes.

Figure 5 The average Energy-Entropy Ratio is calculated for all channel signal, and the scales with higher EE ratio can be selected for constructing the frame cubes (see online version for colours)



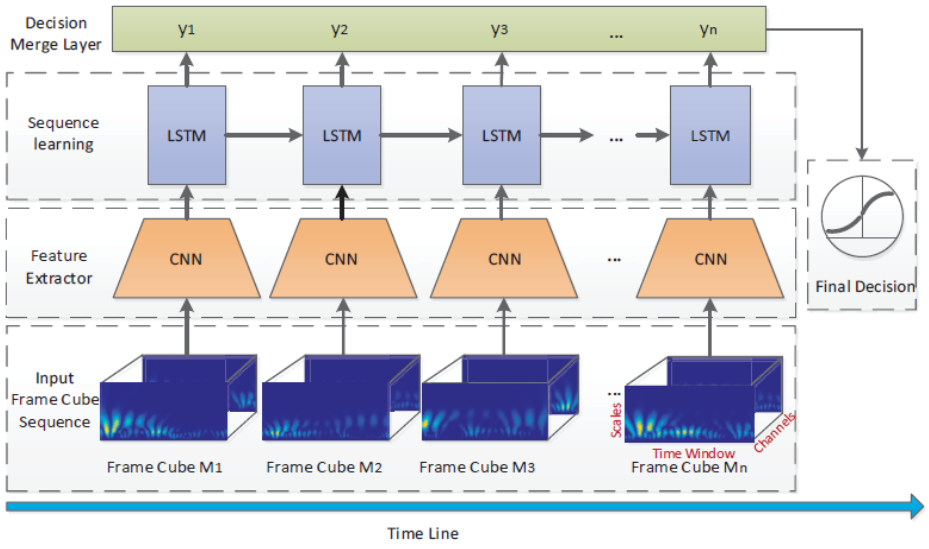
2.3 The convolutional recurrent neural network

Besides the pre-processing method mentioned above, we also propose a hybrid deep learning model, which is called the 'Convolutional Recurrent Neural Networks (C-RNN)' by us, to conduct emotion recognition tasks. As shown in Figure 6, the model is a composition of two kinds of deep learning structures. It combines the powerful ability of the CNN in processing data with grid-like topology and the RNN in processing sequential data. The CNN unit works for mining inter-channel and inter-frequency correlation and extracting features from the frame cubes. The 'Long Short-term Memory (LSTM)' unit, which is a refined RNN structure, models the contextual information for sequences that have arbitrary length. This hybrid model is quite suitable for processing two or three-dimensional sequential data.

The function of the C-RNN in this paper can be formulated as: $< M_1, \ldots, M_t, \ldots, M_n > \leftrightarrow l$, the l represents the predicted emotion label of the sequence. The difference between our model and the traditional 'many-to-one (M2O)' model rests with how the predicted label generated. The label of traditional M2O model is determined by its last step's output, while the label of our model is determined by each step's output.

More specifically, the LSTM based RNN in this model is used for learning contextual information from the feature sequence that extracted through the front CNN, and generating decision information in each time step. The decision merge layer (decision average layer) of the C-RNN is used for recording those decision output, which is the basis for the final decision of the entire trial. Then, the final l is obtained by averaging the value in the softmax nodes of the decision layer in each time step. Then, the node with the maximum average probability determines which class of this trial belongs to, which can be formulated as: $l = argmax\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i\right)$. This strategy complies with our assumption that the participants' emotional rating is based on their entire experience in a trial. We should notice that the weights of the time distributed CNN are tied across time in this model, so it can also be regarded as only a single CNN exists in this time-series model.

Figure 6 The unfolded time chain form of the C-RNN model in emotion recognition. Each time step, a frame cube is fed into the model, the CNN is responsible for extracting inter-channel correlation features through deliberately designed convolutional filters and the extracted information is further fed into the LSTM unit for context learning. Finally the decision layer give the recognition results based on the sequences of the entire trial (see online version for colours)



In summary, the convolutional filter size is deliberately designed for mining the correlation among different channels as well as scales (frequencies). The LSTM based RNN in this model is used for learning contextual information from the feature sequence that extracted through the front CNN, and the emotion recognition for the entire trial is decided based on the output of LSTM in each time step. The model is constructed and trained through some open source deep learning libraries, such as Keras (Chollet, 2015). In the next subsection we will give detailed introduction to the two components CNN and RNN of our hybrid model, respectively.

2.3.1 Convolutional neural networks (CNN)

The Convolutional Neural Networks is a successful case of introducing findings in neuroscience to 'Deep Learning' researches. It has achieved great success not only in the field of computer vision but also in the fields of speech recognition and natural language processing, etc. The architecture and mechanism of CNN provides the possibility for neural networks in processing data with two or three-dimensional structure. The designed convolutional filters help for extracting multiple kinds of features automatically. Generally speaking, a CNN is composed of one or several stacked convolutional layers. Each convolutional layer typically includes three processing stages, namely convolution stage, detector stage and pooling stage (Lecun et al., 1998). The convolutional stage is a process of applying convolutional filters to original 2D data with one or multiple channels in depth. After this process, multiple feature maps are acquired from the input. The characteristics of the convolution stage includes sparse connectivity and parameter sharing. The mechanism of parameter sharing greatly decrease the amount of weight parameters in traditional full-connected neural networks, and in turn reduces the cost for parameter storage. The following detector stage is a non-linear transformation (e.g., ReLU activation function) of the obtained output from prior convolution stage. The last stage is another operation called pooling (e.g., Max Pooling and Average Pooling) which is a summary statistics of nearby results after detector stage, this stage helps the representation to be invariant to translation of input, and meanwhile the size of the input to next convolutional layer or a fully-connected layer can be reduced greatly.

Specifically, in this work the designed convolutional filter of the CNN is used for learning task-related features from the frame cube through mining inter correlation of different channels and different scales in those neurophysiological signals. The characteristics of deep learning in automatic feature extraction, feature selection and feature fusion reduce the difficulties in computing domain-specific features, as well as help us to bypass the traditional feature selection phase. Besides, we can also determine the critical emotion-related neurophysiological variables and components through analysing the trained convolutional filters.

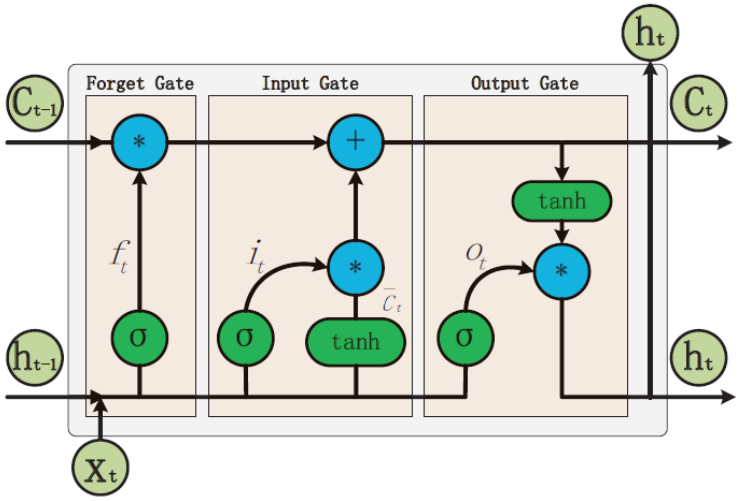
2.3.2 Recurrent neural networks (RNN)

As we know, the magnitudes of physical measurements are quite small, and generally their changes lag behind the evolving of emotions (Krumhansl, 1997). Hence, the RNN is suitable to resolve the delayed effect through accumulating the weak signal characteristics in each time step. After combining with the RNN unit, the hybrid model acquires the ability of learning a time series. The RNN is good at sequential modelling that traditional deep neural network (DNN) can't do well. The difference between the RNN and the DNN is the weights parameters are reused at every time step, so the number of parameters will not increase in proportion with the length of the input sequence. The RNN's ability relies on its recurrence structure, which can model contextual information from sequences with equal or different length. It is very important when we do not know which moment plays the most important role in the subject's final evaluation of the specific emotion they experienced in a trial.

The simple RNN's practical application has been hampered by its special design for difficult training, the RNN must faces the mathematical challenge of 'gradient vanish or explode' in back propagation when its dependencies is too long (Bengio et al., 1994).

Therefore, in order to reduce the difficulties in learning a long-term dependencies, some rectified recurrent units, including GRU and LSTM that combine 'gate' mechanism in their structure, have been adopted to replace the usual units of the traditional RNN. The gate can forget the information has been used and the self-loop structure allows the gradient to flow for long durations. These gated RNNs have gained great success in tasks of handwriting recognition, speech recognition, machine translation, image caption, parsing, etc. (Graves, 2012).

Figure 7 The detailed structure of the LSTM unit, its function for contextual and sequential learning is based on the cooperation of those three gates (see online version for colours)



In this work, we adopt LSTM as the RNN unit. As shown in Figure 6, the recurrence of a LSTM based RNN can be presented and comprehended through unfolding it into a chain form, each chain represents a time step in processing the data that fed from the front CNN. The cell states flow along with those chains, each chain has three gate structures that determine what information from prior step should be forgot and what information in current time step should be added into the main flow. A typical LSTM unit's structure is illustrated in Figure 7, and the mechanism of the gates is described as follows:

The first one is the 'Forget Gate', which determines what information from the past should be forgot. The hidden state h_{t-1} from the prior LSTM cell and the current step's input x_t are concatenated into a new vector, after multiplication with the weight parameters W_f of the gate, each element's value of the output vector f_t is scaled between 0 and 1 through element-wise sigmoidal operation σ. This output f_t acts as a decision vector, it helps to determine what information in the prior cell state C_{t-1} should be reserved through element-wise multiplication: $C_{t-1} * f_t$. The element '0' causes the corresponding information in C_{t-1} will be wiped out, while the element '1' means the corresponding information is allowed passing through. The output f_t of the gate is formalised as equation (8).

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{8}$$

The second one is the 'Input Gate', the fulfilment of its function needs cooperation of two parallel layers. The tangent layer outputs candidate information $\overline{C}_t$ for selection, while the sigmoidal layer acts just as the forget gate, it decides what candidate information will be selected by outputting a decision vector i_t. After the element-wise multiplication of the candidate information by the decision vector $\overline{C}_t * i_t$, the final information that should be added to the cell state is determined. The two layers' functions are formalised as equations (9) and (10), respectively.

$$i_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_i) \tag{9}$$

$$\overline{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t) + b_c) \tag{10}$$

Therefore, the cell state C_t of the current chain is a combination of the reserved historical information of C_{t-1} and the updated information selected from $\overline{C}_t$, as equation (11).

$$C_t = C_{t-1} * f_t + \overline{C}_t * i_t \tag{11}$$

The last one is the Output Gate'. In a word, it decides outputting what hidden state h_t in current chain through multiplication of the decision vector o_t by the candidate information selected from C_t, as shown in equations (12) and (13), respectively.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{12}$$

$$h_t = \tanh(C_t) * o_t \tag{13}$$

3 Experiment and discussion

In this section, we will show the effectiveness of our methods, and we also compare our method with several classical baseline methods.

3.1 Experimental dataset

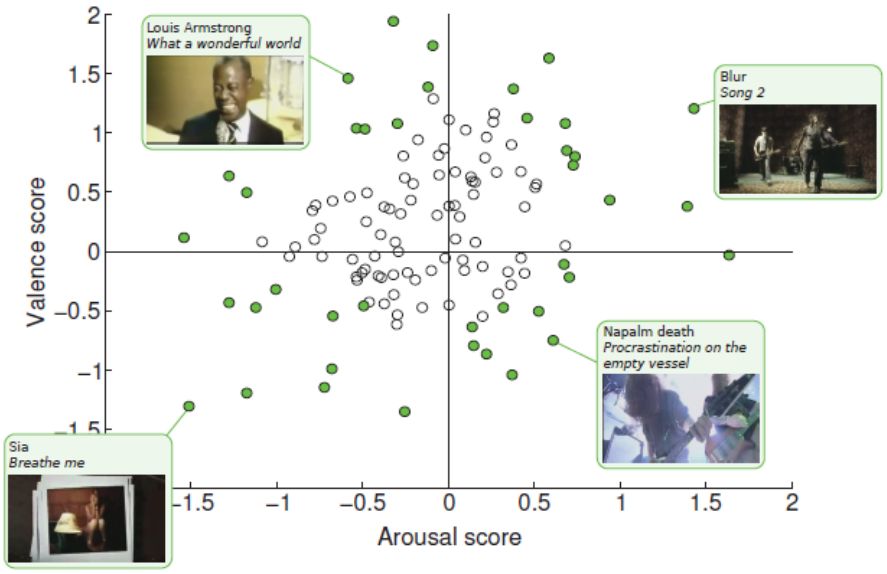
Our model was validated on the target dataset DEAP (Koelstra et al., 2012), which includes multi-channel neurophysiological signals collected from 32 subjects. The subjects' various emotions were stimulated through 40 one-minute-long music videos that corresponding to different emotional genres. One stimuli is presented in one trial, and the signals were continually recorded during those trials. The 40 stimuli are selected from 120 candidate music videos based on some volunteers' ratings, which is projected on a two-dimensional emotional space proposed by Russell (Russell, 1980), as illustrated in Figure 8, where the two dimensions are Arousal (it ranges from relaxed to aroused) and Valence (it ranges from pleasant to unpleasant), respectively. The higher the specific ratings the stronger the specific emotion that the video contains.

3.2 Label processing

Quantifying and representing the emotional experience is the precondition for computing and recognising emotions. Besides the six discrete emotional states proposed by Paul Ekman (Ekman, 1993), which is firstly adopted in facial expression based studies. The

'Valence-Arousal Plane' based two-dimension-space representation method is currently being widely adopted, as there are fuzzy boundaries existing in the transition of various emotional experiences. The emotional dimension of 'Valence' is a measure of the degree of happiness or sadness the subjects feel, and another dimension of 'Arousal' reflects the intensity of excitement. All kinds of emotional states can be projected on the 2D plane with a value between 1 to 9 for 'Valence' and 'Arousal', respectively. In this study, according to the subjects' personal ratings, we divide and label the trials into two classes for Valence and Arousal respectively (pleasant: > 5, unpleasant: ≤ 5; aroused: > 5, relaxed: ≤ 5).

Figure 8 The ratings on Valence-Arousal plance for 120 candidate music videos. The 40 videos that lie closest to the extreme corners of the four quadrants are selected as the stimuli materials (see online version for colours)

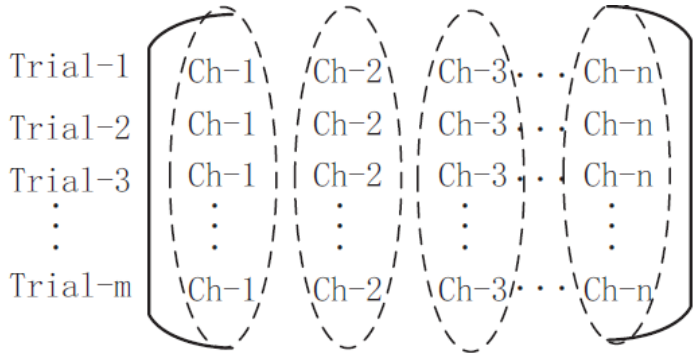


3.3 *Data processing*

For comparing our approach with the traditional feature engineering based methods, we need determining and extracting some hand-craft features.

Data Normalisation: Before extracting features, we conduct data normalisation by scaling each channel's signal as illustrated in Figure 9. The data of each subject are normalised per channel for all the trials. This procedure helps removing subject bias and generating more comparable features between subjects. Meanwhile, the variability of different channels can be preserved.

Figure 9 The channel data normalisation paradigm adopted for baseline methods



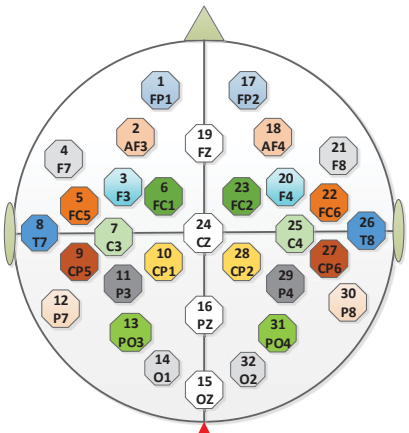
Feature Extraction: The EEG is a kind of reflection of the brain rhythms. Traditionally, we divide the brain rhythms into *Delta* rhythm (<3Hz), *Theta* rhythm (4Hz-7Hz), *Alpha* rhythm (8Hz-12Hz), Beta rhythm (13Hz-30Hz) and Gamma rhythm (>31Hz) according to different frequency components consisted in the EEG. Various neuropsychological studies try to correlate those different rhythms with the underling brain functions and cognitive processes. Therefore, in this work we filtered out the four different rhythms from the scaled EEG signal, and conducted feature extraction for each of the four extracted rhythms. The concatenation of the four vectors of rhythm features is adopted as the training samples for baseline methods.

The features that we compute and extract for EEG can be summarised into three main categories, including time-frequency-domain features, nonlinear-dynamics features and brain-asymmetry features. For the time-frequency domain, we extracted 9 different features. We also extracted 9 different nonlinear-dynamics features, as researchers found the brain manifests many characteristics specifically belongs to chaotic dynamical systems (Stam, 2005, 2006). Brain asymmetry oriented neuropsychological studies have emphasised the mediator or moderator role of those rhythms in emotional information processing and emotional responses (Davidson, 1992; Coan and Allen, 2004; Mathersul et al., 2008). Hence, besides the above mentioned features, considering the asynchronous activity phenomenon in two hemispheres, we also extracted 14 brain-asymmetry features for symmetrically located electrode pairs, which we called brain-asymmetry features. The total kinds of handcraft features extracted for EEG is illustrated in Figure 10.

The target four kinds of rhythm components are extracted through designed 'Finite Impulse Response (FIR)' bandpass filter with 'Hanning window'. After that, we calculated the above mentioned handcraft features for each of the four rhythms with a 4s sliding window and 2s overlap, then the average of the results calculated and obtained in those sliding windows is adopted as the feature. For 32-channel EEG, the overall number of features extracted for each rhythm is: $(9 + 9) \times 32 + 14 = 590$, hence the dimension of the feature vector for one trid is $590 \times 4 = 2360$. The extracted handcraft features and the implementation of the C-RNN model can be found on the website: https://github.com/muzixiang/Multichannel_Biosignal_Emotion_Recognition.

Figure 10 The selected three kinds of features computed for the baseline methods. The right part of the figure illustrates the distribution of the 14 symmetrically located electrode pairs (see online version for colours)

Feature Types	Feature Details
Time-Frequency Feature (9 features x 32 channels x 4 bands)	the peak-to-peak mean value, the mean square value, the variance, the maximum power spectral density, and its corresponding frequency, the sum power, the three Hjorth parameters (the activity, the mobility, and the complexity)
Nonlinear Dynamic Feature (9 features x 32 channels x 4 bands)	the approximate entropy, the c0 complexity, the correlation dimension, the Lyapunov exponent, the Kolmogorov entropy, the permutation entropy, the singular entropy, the Shannon entropy, the spectral entropy
Brain Asymmetry Feature (14 features x 4 bands)	FP1-FP2, AF3-AF4, F3-F4, F7-F8, FC5-FC6, FC1-FC2, C3-C4, T7-T8, CP5-CP6, CP1-CP2, P3-P4, P7-P8, PO3-PO4, O1-O2.



3.4 *Experimental settings*

In the experiment, we build subject-dependent model for each subject, and we adopt the leave-one-trial-out strategy to validate the performance. More specifically, in each iteration, one trial of the subject are left and the other trials are used as training data. After training, the model gives a prediction on the left test trial. This process iterates until the subject's last trial has been predicted. Then the performance can be measured through comparing the gathered predictions with the trials' original labels. In the experiment, the F-score is adopted as the performance metric to evaluate and compare the different method. The F-score distributions of each method on all 32 subjects are illustrated in a box plot manner.

Baseline Method: For comparison, we take into account several representative classification methods that belongs to different 'Machine Learning' categories. The selected methods including Gaussian Naive Bayes (NB) that based on Bayes' theorem, Support Vector Machine (SVM) that based on statistical learning theories (c = 1.0; *kernel = linear*), Logistic Regression (LR) that based on linear function (*penalty = 12; solver = liblinear*), K Nearest Neighbour (KNN) that based on distance metric (*n_neighbours* = 5), Gradient Boosting Decision Tree (GBDT) that based on ensemble learning theory (*n_estimators* = 100; *criterion = friedman_mse*) and Multi-layer Perceptron (MLP) that based on artificial neural network (*hidden_layer_size* = {100,10}; *alpha* = 0.0001; *activation = relu; solver = Ibfgs; learning_rate* = 0.001). Those methods are

implemented by the Scikit-learn toolkit (Pedregosa et al., 2012), and most of the parameters are set to their default values.

C-RNN Model: The structure and the detailed settings for hyper parameters of the C-RNN model is illustrated in Figure 11. In this work, we adopt two stacked convolutional layers as the basic structure of the time-distributed CNN, the size of the convolutional filter in the first convolutional layer is specifically set as 1×1 with the purpose of mining and fusing inter-channel correlation information in each scale and time point. In this stage, we set the number of the convolutional filters as 16, after repeated convolutional operation, we obtain 16 feature maps that contains different kinds of mined inter-channel correlation information, each remains the original size in dimension of scale and time. Following the first convolutional operation is another convolutional operation, where we set the number of kernel as 32 and the size as 32×1, for further fusing and correlating the inter-scale information based on the input 16 feature maps. After the two convolutional stages, the original frame cube has been transformed into 32 one-dimensional feature maps. Afterwards, a subsampling operation called average pooling was applied to the 32 one-dimensional feature maps, the pooling size is 1×8 for aggregating the redundant information in adjacent 8 time points. After this pooling stage, the size of each feature map is down sampled from 1×128 into 1×16. Before feeding those feature maps into the LSTM unit, a operation called flatten is needed for transforming and concatenating the different feature maps into a single one-dimensional vector. After the processing in LSTM unit, there are two time-distributed fully connected layers for processing the LSTM's output in each time step. The second fully connected layer is the decision layer with softmax function, the number of its nodes is identical to the number of emotional classes we hope to recognise. The final decision is made in the decision merge operation through averaging the probability of decision nodes in each time step, the node with the highest average probability decides the class that the trial belongs to. We can adopt 'dropout' in full-connected layers to prevent over-fitting (Srivastava et al., 2014). In the experiment, the length of the sliding window in constructing the frame cubes for the C-RNN model is set as one-second long.

Figure 11 The settings for the C-RNN in processing multi-channel EEG frame Cubes and conducting emotion recognition

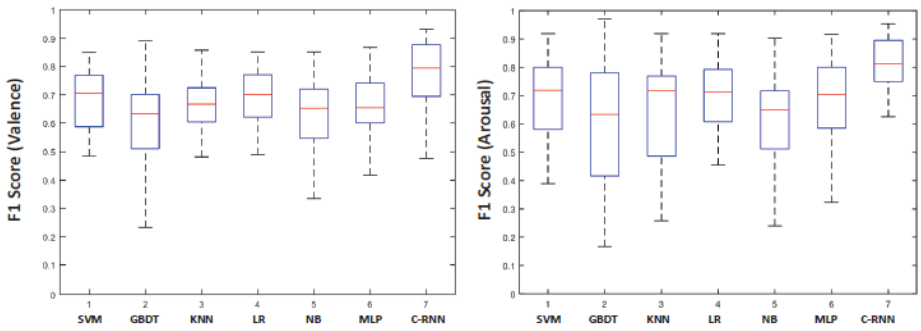
Input Data Shape	Operation	Setting	Discription
Frame Cube Seqence: 32@32x128	TimeDistributedConvolution2D +Relu	16@1x1	Fusing and correlating the cross-channel information.
16@32x128	TimeDistributedConvolution2D +Relu	32@32x1	Fusing and correlating the cross-scale information.
32@1x128	TimeDistributedAveragePooling2D	1x8	Fusing information for adjacent time points.
32@1x16	TimeDistributedFlatten	None	Flattening the obtained feature maps into one-dimensional vector.
1x512	LSTM	1x128	Connecting the flattened one-dimensional vector to LSTM for context learning.
1x128	TimeDistributedDenseLayer	1x64	Connecting each time step's output of the LSTM to a fully connected layer.
1x64	TimeDistributedDenseLayer	1x2	Further fully connecting to a decision layer with softmax function.
Decision Sequence: 1x2	TimeDistributedDecisionMerge	1x2	Averaging each time step's decision for final decision making.

3.5 Results and discussion

The F-score of each subject obtained through the baseline methods and the C-RNN model are illustrated in a box plot manner in Figure 12. Overall, as we can see, the C-RNN model greatly outperforms the baseline methods with a highest mean F-score (0.7192 on Valence and 0.7742 on Arousal) and a relatively low standard deviation (0.1593 on Valence and 0.1257 on Arousal). The performance on Arousal is better than that on Valence, perhaps because the degree of Arousal is more like an indicator of neural activations, which can be reflected well by the wavelet energy spectral, whereas, compared to Arousal, the recognition of Valence is a more complicated task and cannot be explicitly differentiated only through analysing the fluctuation of signal energy. It is also possible that the experience on Valence is intrinsic fuzzy, those subjects cannot clearly determine and rate the degree of Valence, which introduces some noise into the labels. Anyway, the C-RNN model combined with spectral based frame cube representation achieves satisfied and reasonable results on emotion recognition, with respect to the dimensions of Valence and Arousal.

Prior to this work, most studies focus on segment-level emotion recognition, whereas, for trial-level emotion recognition, as we know there exists only a few of relevant works. For example, Chen et al. (2015) extracted over a thousand of features and studied different feature selection method and adopted Hidden Markov Models (HMM) to perform trial-level emotion recognition on a subset of the total 32 participants, and finally obtained nearly 73.0% and 75.6% mean classification accuracy (MCA) on Valence and Arousal, respectively. Rozgic et al. (2013) proposed four different fusion strategies based on segment-level features for constructing trial-level features, and conducted recognition through K-PCA and RBF-SVM for the trial, the highest MCA reported on Valence and Arousal is 76.9% and 68.4%, respectively. Koelstra et al. (2012) simply extracted several trial-level features and performed recognition through Naive Bayes classifier, the results are not better than our proposed baselines in this work.

Figure 12 Comparison of the recognition performance on 32 subjects between different methods



Besides the effectiveness of our approach in emotion recognition, another advantage of our approach is it greatly reduce the time cost in preparing the data and features. As shown in Figure 13, which illustrates the time cost in pre-processing and preparing one subject's samples (processing through a PC with i7-4770 CPU and 8 Gb Ram). The frame cube sequence construction (Method 2) is almost 40 times faster than the traditional feature engineering based pre-processing method (Method 1). Empirical

analysis shows that most of the time cost in feature engineering based approaches rests with computing and extracting various non-linear features. Considering that the time cost in Method 1 that presented here only refers to one kind of frequency band, if we conduct pre-processing for band of Theta, Alpha, Beta, Gamma in a serial way and if the signal length is even longer, the time cost must be higher. Most importantly, in clinical application, we cannot wait such long for the results, meanwhile, currently the portable equipment's' computational ability cannot support such a high burden processing.

Figure 13 The time cost in preparing the data for one subject. The first bar represents the feature engineering based method, whereas the second bar represents the proposed frame cube based method

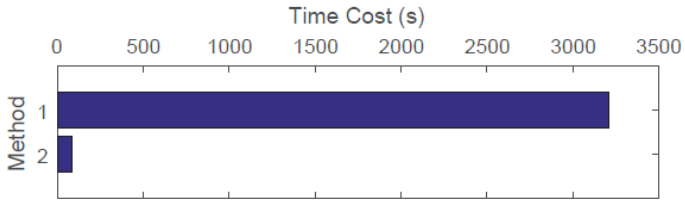
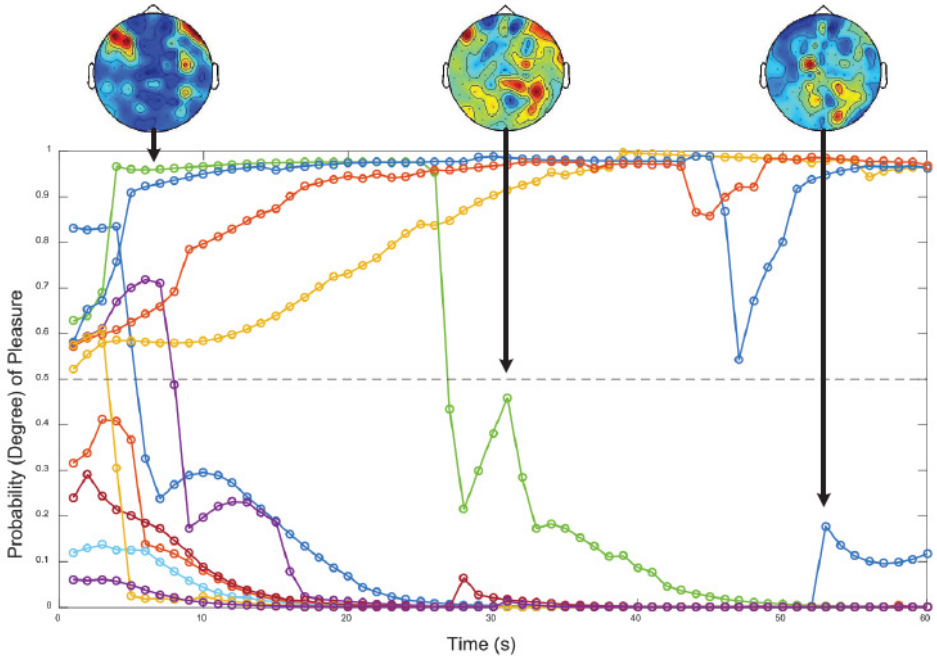


Figure 14 The real-time prediction on Valence for several 60s long trials of one subject. The results are obtained through storing each time step's decision probability. The probability greater than 0.5 indicates a pleasant experience, while the probability lower than 0.5 indicates an relatively unpleasant experience. The brain activity maps are also annotated for several key time points in a trial (see online version for colours)



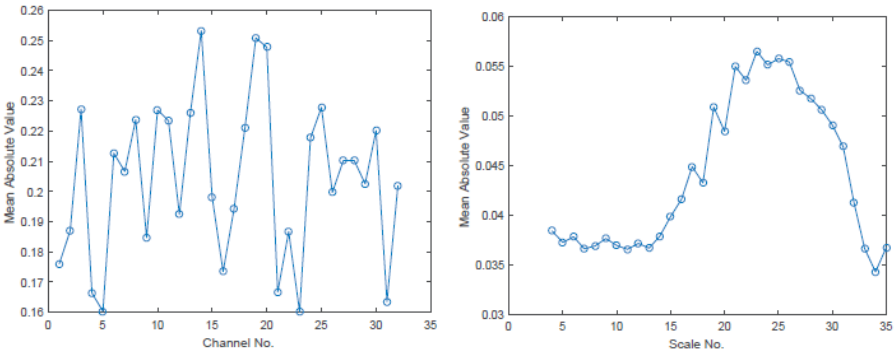
As shown in Figure 14, the RNN based model not only performs well in recognising subjects' emotional experience in the entire trial, but also it has the ability in real-time prediction for each time step. On the premise of real time prediction, if we can map the brain activity of each time point with its corresponding emotional states, we may find a common pattern in emotional cognition process. However, the DEAP dataset does not provide us with real-time labels for each time step, we could not validate the accuracy of the output predictions currently. Even so, there indeed exists an urgent need in real-time emotion monitoring for applications in clinical psychiatry and emotional brain-computer interaction (BCI), in which the system could log the emotional transitions in detail and give quick responses or warnings to unexpected emotional fluctuations.

This model also could give us a new perspective in studying critical brain areas in emotional cognition process, as well as could provide us with a new feature selection method. Based on the trained C-RNN model, we can determine which channel and scale is critical in reflecting various emotional process. In other words, the designed convolutional filters automatically perform channel selection and scale selection in the training process, which is quite different with the traditional feature selection method in feature engineering. Specifically, the convolutional kernel in the first time-distributed convolutional layer is set as 1×1 to learn inter-channel information, meanwhile the importance of each channel and corresponding brain area can be reflected through analysing corresponding weights. The higher the absolute value of the weight denotes a more contribution of its information in reflecting specific emotional states. Similarly, the kernels in the second time-distributed convolutional layer are set as $S \times 1$ to learn inter-scale information, it helps us determine which scale contributes most in the task, as well as helps us determine if there exists correlation between specific EEG component and specific emotional process.

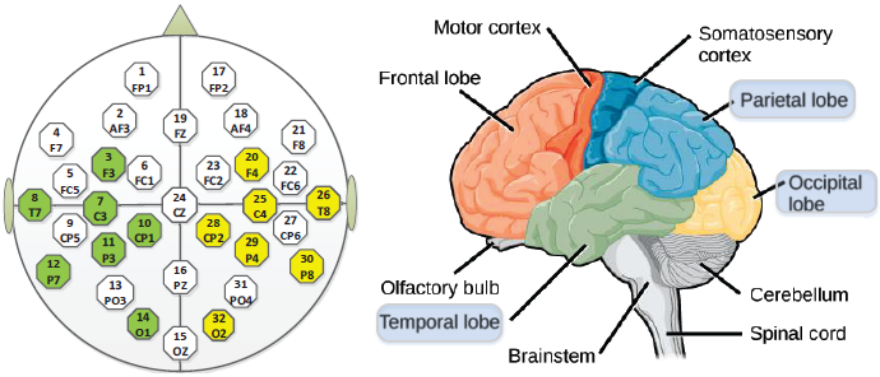
Deriving from the Figure 15(a), we can rank the weight value (importance) of those channels in ascending order, as follows: FC2, FC5, PO4, F7, F8, PZ, FP1, CP5, FC6, AF3, P7, FP2, OZ, T8, O2, P4, C3, CP6, CP2, FC1, CZ, P8, AF4, P3, T7, PO3, CP1, F3, C4, F4, FZ, O1. Based on the obtained importance of each single channel, we further sort them as symmetrical pairs in descending order, namely F3-F4, O1-O2, CP1-CP2, C3-C4, P3-P4, T7-T8, P7-P8, AF3-AF4, CP5-CP6, PO3-PO4, FC1-FC2, FP1-FP2, FC5-FC6, F7-F8. The top 50% critical pairs are annotated in Figure 15(c), as we can see they mainly distribute in brain's parietal-temporal-occipital (PTO) region, which is in line with the findings in some neuropsychological works (such as Adolphs, 2002; Grecucci et al., 2013; Sarkheil et al., 2013). These studies provide evidence that the parietal lobe, temporal-parietal junction (TPJ) and occipital lobe are correlated in various emotional information processing. We also sum the weights of channels in left and right brain hemisphere respectively, and we find that the total weights of the right hemisphere is slightly greater than that of left hemisphere, it is also in line with the statements of some neuropsychological studies that the right hemisphere plays critical roles in emotional information processing (Adolphs, 2002). As illustrated in Figure 15(b), the scales with relatively higher weight roughly distribute from scale-15 to scale-32, whose corresponding frequency locates within the scope of *Theta* rhythm, roughly from 3Hz to 8Hz. This finding does not agree with some neuroscience studies that claim emotion related information mainly exists in higher frequency bands, such as *Beta* band and *Gamma* band (Muller et al., 1999; Kortelainen et al., 2015). Nevertheless, it also should be noted that some other neuroscience studies support our finding, they find low-frequency rhythm also correlates with emotional information processing (Knyazev et al.,

2009; Uusberg et al., 2014). Hence, in next work we need further studying on more subjects, and determining whether it is a common phenomenon or just one exception.

Figure 15 The mean absolute value (MAV) of the weight for each channel (a) and each scale (b) derived from one subject's model. The higher the MAV the more critical the corresponding component in emotion recognition. The top 50% critical channel pairs are also analysed and annotated in the topology of the 10-20 brain system (c) (see online version for colours)



(a) Weights for Channel (b) Weights for Scale



(c) Distribution of Critical Channel Pairs

In summary, compared to those relevant methods, the C-RNN model performs better and is a good choice when the physiological signal is long and the contextual physical information is needed. The CNN based feature learning strategy only needs simple data pre-processing, which is time saving and easy for non-experts to master, meanwhile, it provides us with a new approach in determining critical emotion-related rhythms and brain regions through analysing the learned kernel weights. Besides, this model is suitable for clinical applications that need incremental learning (the samples are continuously gathered and fed into the model) and real-time emotion monitoring.

4 Conclusions

In this paper, we have proposed a hybrid deep learning model, C-RNN, which integrates CNN and RNN, for emotion recognition and monitoring based on multi-channel EEG signals. Specifically, the CNN component has the ability in fusing, mining and selecting inter-channel and inter-frequency correlation information. On the other hand, the RNN (i.e., LSTM) based model structure can learn long-term dependencies and contextual information from the constructed frame cube sequences. Practically, instead of manually designing task related features as in traditional approaches, we propose a novel pre-processing method that transforms the multi-channel EEG data into a 3D frame cube representation, which greatly reduces the time cost in data pre-processing, compared with traditional feature extraction and selection paradigm. The proposed method outperforms classical feature engineering based machine learning methods in the trial-level emotion recognition task. The analysis on the trained convolutional filters can give us another perspective of decoding the brain's cognitive scheme in affective information processing, meanwhile, benefits traditional feature engineering approaches, e.g., it will spend less time in data pre-processing through extracting features only from the critical channels and frequencies. It also has a potential in giving predictions not only for an entire trial but also for each time point, which is very important in real-time emotion monitoring scenarios. The potential in real time prediction provides a new way in mapping the brain activity of specific time with a specific emotional state, through which we may find a common pattern in emotional cognition process. There could be room in improving the performance of our model if we can augment the training data and reduce the influence of individual difference, further, a subject-independent model is needed for public usage. All these problems may be resolved through adopting various transfer learning strategies that we will investigate in future works.

Acknowledgements

This work is funded in part by the Chinese National Program on Key Basic Research Project (973 Program, grant no. 2014CB744604, 2013CB329303 and 2013CB329304), Chinese 863 Program (grant no. 2015AA015403), Natural Science Foundation of China (grant no. U1636203 and 61402324), and Tianjin Research Program of Application Foundation and Advanced Technology (grant no. 15JCQNJC41700).

References

Adolphs, R. (2002) 'Neural systems for recognizing emotion', *Current Opinion in Neurobiology*, Vol. 12, No. 2, pp.169–177.

American Psychiatric Association (2013) 'Diagnostic and statistical manual of mental disorders (DSM-5®)', American Psychiatric Association, Arlington.

Bengio, Y., Simard, P. and Frasconi, P. (1994) 'Learning long-term dependencies with gradient descent is difficult', *IEEE Transactions on Neural Networks*, Vol. 5, No. 2, pp.157-166.

Blanchette, I. and Richards, A. (2010) 'The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning', *Cognition and Emotion*, Vol. 24, No. 4, pp.561–595.

Bolos, V. J. and Benitez, R. (2014) 'The wavelet scalogram in the study of time series', *Advances in Differential Equations and Applications*, Vol. 4, No. 3, pp.147–154.

Chen, J., Hu, B., Xu, L., Moore, P. and Su, Y. (2015) 'Feature-level fusion of multimodal physiological signals for emotion recognition', *Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.395–399.

Chollet, F. (2015) 'Keras: Deep learning library for theano and tensorflow', https://github.com/fchollet/keras.

Coan, J. A. and Allen, J.J.B. (2004) 'Frontal EEG asymmetry as a moderator and mediator of emotion', *Biological Psychology*, Vol. 67, Nos. 1–2, pp.7–50.

Davidson, R. J. (1992) 'Anterior cerebral asymmetry and the nature of emotion', *Brain and Cognition*, Vol. 20, No. 1, pp.125–151.

Duan, R., Zhu, J. and Lu, B. (2013) 'Differential entropy feature for EEG-based emotion classification', *Proceedings of the 6th IEEE/EMBS International Conference on Neural Engineering*, pp.81-84.

Ekman, P. (1993) 'Facial expression and emotion', *American Psychologist*, Vol. 48, No. 4, pp.384–392.

Ekman, P. and Davidson, R. J. (1994) 'The nature of emotion', *American Scientist*, Vol. 6, No. 1, pp.3–31.

Gandhi, T, Panigrahi, B. K. and Anand, S. (2011) 'A comparative study of wavelet families for EEG signal classification', *Neurocomputing*, Vol. 74, No. 17, pp.3051–3057.

Graves, A. (2012) 'Supervised sequence labelling with recurrent neural networks', *Studies in Computational Intelligence*, Vol. 385, pp.5–13.

Grecucci, A., Giorgetta, C., Bonini, N. and Sanfey, A. G. (2013) 'Reappraising social emotions: The role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation', *Frontiers in Human Neuroscience*, Vol. 7, pp.523.

Hernandez, J., McDuff, D., Benavides, X., Amores, J., Maes, P. and Picard, R. (2014) 'AutoEmotive: Bringing empathy to the driving experience to manage stress', *Proceedings of the 2014 Companion Publication on Designing Interactive Systems*, pp.53–56.

Huang, D., Guan, C., Ang, K., Zhang, H. and Pan, Y. (2012) 'Asymmetric spatial pattern for EEG-based emotion detection', *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*, pp.1–7.

Hussain, M. S., Calvo, R.A. and Pour, P.A. (2011) 'Hybrid fusion approach for detecting affects from multichannel physiology', *Proceedings of the 2011 International Conference on Affective Computing and Intelligent Interaction*, pp.568–577.

Jirayucharoensak, S., Pan-Ngum, S. and Israsena, P. (2014) 'EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation', *The Scientific World Journal*, Vol. 2014, Article ID. 627892.

Kim, H. S., Eykholt, R. and Salas, J. D. (1999) 'Nonlinear dynamics, delay times, and embedding windows', *Physica D Nonlinear Phenomena*, Vol. 127, Nos. 1–2, pp.48–60.

Knyazev, G. G., Slobodskoj-Plusnin, J.Y., and Bocharov, A.V. (2009) 'Event-related delta and theta synchronization during explicit and implicit emotion processing', *Neuroscience*, Vol. 164, No. 4, pp.1588–1600.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I. (2012) 'Deap: A database for emotion analysis using physiological signals', *IEEE Transactions on Affective Computing*, Vol. 3, No. 1, pp.18–31.

Kortelainen, J., Yrynen, E., Sepp, Nen, T. (2015) 'High-frequency electroencephalographic activity in left temporal area is associated with pleasant emotion induced by video clips', *Computational Intelligence & Neuroscience*, Vol. 2015, No. 6, pp.762–769.

Krumhansl, C. L. (1997) 'An exploratory study of musical emotions and psychophysiology', *Canadian Journal of Experimental Psychology*, Vol. 5, No. 4, pp.336–353.

Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, Vol. 86, No. 11, pp.2278–2324.

Li, Mu and Lu, Baoliang (2009) 'Emotion classification based on gamma-band EEG', *Proceedings of the 2009 International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pp.1223–1226.

Li, X., Zhang, P., Song, D. and Hou, Y. (2015) 'Recognizing emotions based on multimodal neurophysiological signals', *Advances in Computational Psychophysiology*, pp.28–30.

Li, X., Hu, B., Sun, S. and Cai, H. (2016) 'EEG-based mild depressive detection using feature selection methods and classifiers', *Computer Methods and Programs in Biomedicine*, Vol. 136, pp.151–161.

Liu, C., Agrawal, P., Sarkar, N. and Chen, S. (2009) 'Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback', *International Journal of Human-Computer Interaction*, Vol. 25, No. 6, pp.506-529.

Lopatovska, I. and Arapakis, I. (2011) 'Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction', *Information Processing and Management*, Vol. 47, No. 4, pp.575–592.

Louis, A. S. and Laurel, J. T. (2001) 'Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions', *Cognition and Emotion*, Vol. 15, No. 4, pp.487–500.

Mathersul, D., Williams, L. M., Hopkinson, P. J. and Kemp, A. H. (2008) 'Investigating models of affect: Relationships among EEG alpha asymmetry, depression, and anxiety', *Emotion*, Vol. 8, No. 4, pp.560-572.

Minsky, M. (1988) 'The society of mind.', Simon & Schuster, New York.

Muller, M. M., Keil, A., Gruber, T. and Elbert, T. (1999) 'Processings of affective pictures modulates right-hemispheric gamma band EEG activity', *Clinical Neurophysiology Official Journal of the International Federation of Clinical Neurophysiology*, Vol. 110, No. 11, pp.1913-1920.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2012) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, Vol. 12, No. 10, pp.2825–2830.

Picard, R. W. (1999) 'Affective Computing for HCI', *Proceedings of the 8th International Conference on Human-Computer Interaction: Ergonomics and User Interfaces*, Vol. 1, pp.829-833.

Rozgic, V., Vitaladevuni, S. N., and Prasad, R. (2013) 'Robust EEG emotion classification using segment level decision fusion', *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1286–1290.

Rubinstein, R., Bruckstein, A. M. and Elad, M. (2010) 'Dictionaries for sparse representation modeling', *Proceedings of the IEEE*, Vol. 98, No. 6, pp.1045–1057.

Russell, J. A. (1980) 'A circumplex model of affect', *Journal of Personality and Social Psychology*, Vol. 39, No. 6, pp.1161–1178.

Sarkheil, P., Goebel, R., Schneider, F. and Mathiak, K. (2013) 'Emotion unfolded by motion: A role for parietal lobe in decoding dynamic facial expressions', *Social Cognitive Affective Neuroscience*, Vol. 8, No. 8, pp.950–957.

Shen, L., Wang, M. and Shen, R. (2009) 'Affective e-Learning: Using emotional data to improve learning in pervasive learning environment', *Educational Technology & Society*, Vol. 12, No. 2, pp.176–189.

Srivastava, N., Hinton, G. E. and Krizhevsky, A. and Sutskever, I. and Salakhutdinov, R. (2014) 'Dropout: A simple way to prevent neural networks from overfitting', *Journal of Machine Learning Research*, Vol. 15, No. 1, pp.1929–1958.

Stam, C. J. (2005) 'Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field', *Clinical Neurophysiology*, Vol. 116, No. 10, pp.2266–2301.

Stam, C. J. (2006) 'Nonlinear brain dynamics', Nova Science Publishers, New York. Subasi, A. (2005) 'Automatic recognition of alertness level from EEG by using neural network and wavelet coefficients', *Expert Systems with Applications*, Vol. 28, No. 4, pp.701–711.

Uusberg, A., Thiruchselvam, R. and Gross, J. J. (2014) 'Using distraction to regulate emotion: Insights from EEG theta dynamics', *International Journal of Psychophysiology*, Vol. 91, No. 3, pp.254–260.

Ward, L. M. (2003) 'Synchronous neural oscillations and cognitive processes', *Trends in cognitive sciences*, Vol. 7, No. 12, pp.553–559.

World Health Organization (2001) 'Mental health: A call for action by world health ministers', Department of Mental Health and Substance Dependence, World Health Organization, Geneva.

Yan, H. and Pham, T (2006) 'Spectral similarity for analysis of DNA microarray time-series data', *International Journal of Data Mining and Bioinformatics*, Vol. 1, No. 2, pp.150-161.

Zheng, F., Shen, X., Fu, Z., Zheng, S. and Li, G. (2010) 'Feature selection for genomic data sets through feature clustering', *International Journal of Data Mining and Bioinformatics*, Vol. 4, No. 2, pp.228–240.

Zheng, W., Zhu, J. and Lu, B. (2016) 'Identifying stable patterns over time for emotion recognition from EEG', *ArXiv Preprint*, arXiv:1601.02197.