This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

EEG-based emotion recognition using hierarchical network with subnetwork nodes

Yimin Yang, *Member, IEEE,* Q. M. Jonathan Wu, *Senior Member, IEEE,* Wei-Long Zheng, *Student Member, IEEE* and Bao-Liang Lu, *Senior Member, IEEE*

***Abstract*—Emotions play a crucial role in decision-making, brain activity, human cognition, and social intercourse. This paper proposes a hierarchical network structure with subnetwork nodes to discriminate three human emotions: positive, neutral, and negative. Each subnetwork node embedded in the network that are formed by hundreds of hidden nodes, could be functional as an independent hidden layer for feature representation. The top layer of the hierarchical network, like the mammal cortex in the brain, combine such features generated from subnetwork nodes, but simultaneously, recast these features into a mapping space so that the network can be performed to produce more reliable cognition. The proposed method is compared with other state-of-the-art methods. The experimental results from two different EEG datasets show that a promising result is obtained when using the proposed method with both single and multiple modality.**

***Index Terms*—Electroencephalogram (EEG), Feedforward neural network, Subnetwork nodes, Emotion recognition.**

I. Introduction

Brain activity for recognition and control has been well-established for several decades. Recently, extraction of additional brain information regarding the psychological states from neurophysiological signals has earned an increased amount of attention in the human-machine-interaction field. To make human-machine-interaction more natural, comprehend about human emotional state is considered as an important factor. Most of the measures utilized to observe physiological states are "non-invasive", based on collecting signals from different modalities (e.g., face, motion, eye, brain, posture, and skin). Among the various methods to emotion recognition, electroencephalography (EEG)-signals based algorithms are being increasingly used due to its high accuracy and stabilization [1][2]. Early work on EEG-based emotion recognition dates back as far as 1985 [3][4][5][6][7]. Intelligence computational approaches from the field of machine learning are widely used to boost recognition performance, which has gained more and more attentions. [8], [9], [10], [11], [12], [13].

First comes the most prominent methods that utilize statistical-based, wavelet-based, fusion-based algorithms for EEG-based feature processing. After this, classification methods are provided, such as support vector machine (SVM), fuzzy *k*-means, single layer feedforward network (SLFN), which resulted in moderate emotion recognition percentages for up to two[14], three[15], four [16], and five emotion states. For example, Lin *et al.* [16] adopt the F-score index which is based on the ratio of between-class and within-class emotion recognition. They gained an average of 82.29% classification accuracy for four emotions across 26 subjects/participants. Chanel *et al.*[17] reported an accuracy of 63% for three emotion states using EEG time-frequency feature. Furthermore, by fusion of the different features and rejection of non-confident samples, they finally obtained an average of 80% classification accuracy. Zheng and Lu [18] proposed selecting 12 channel electrodes features in SVM where these features were preprocessed by a differential entropy (DE) method [19], and then, a LIBSVM was utilized for classification. They showed that exactly 12 electrodes orders with SVM could provide a relative stability with the best accuracy of 86.65%, which outperformed the result of full 62 electrodes. Furthermore, for multimodal emotional recognition, researchers adopted both the feature level [20][2][21] and the decision level fusion [22][23]. Takahashi [24] indicated an emotion recognition method using multiple modality signals (EEG, pulse, electromyogram (EMG), electrocardiogram (ECG) and skin resistance). Zheng *et al.* [25] indicated a fusion-based emotion recognition method by using the multiple modality signals (eye movement and EEG), which showed the recognition rate increased from 76% to 87%.

Another leading trend for deep learning (DL) based emotion recognition. DL has been around for many years, dating back to the works in the 1980s [26], [27], [28], [29], [30], [31]. The *Neocognitron* [28] could be the first artificial neural network that deserved the attribute "deep", and was the first to incorporate neurophysiological insights. In 2006, Hinton [32] initiated a breakthrough in feature extraction, which was quickly followed up in successive years [33], [34], [35], [36]. Various studies [32][34][37][38][39] showed that multilayer neural networks (NNs) with iteration methods or non-iteration methods can be used for representation learning. Powered by the novel method, DL-based learning methods penetrated into EEG emotional recognition field.

Y. M. Yang is with the Department of Electrical and Computer Engineering, University of Windsor N9B 3P4, Canada.

Q. M. Jonathan Wu is with the Department of Electrical and Computer Engineering, University of Windsor N9B 3P4, Canada, and is also with the Department of Computer Science and Engineering, Shanghai Jiao Tong University and the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (Corresponding Author: jwu@uwindsor.ca)

W. L. Zheng and B. L. Lu are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University and the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, and Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai 200240, China

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

Martinez *et al.* employed several convolutional layers in order to learn to obtain the relevant features from the two physiological signals individually for discriminating the four emotion states (relaxation, anxiety, excitement, and fun). Zheng *et al.* [40] trained a deep belief network (DBN) with differential entropy (DE) features and achieved 87.62% classification accuracy.

However, some problems still remain. In fact, the human emotion generation involved in understanding the situation can be a complicated and subjective process. Emotions reflect the biological cognitive processes associated with biological understanding and psychophysiological phenomena, and thus, it is difficult to propose a recognition method which is purely based on traditional machine learning methods. For example, according to recent studies, the thalamus, basal ganglia, insular cortex, amygdala, and frontal cortex are all involved in emotion recognition [41]. Furthermore, accumulated direct biological evidence [42][43] supports the theory that neuron activity in a mammal's prefrontal cortex is heterogeneous, partially random, and disordered. Crucially, the combined features extracted from mixed selectivity neurons may be central to complex cognition. Motivated by these biological evidences, this paper proposes novel hierarchical network methods for EEG-based emotion recognition. In particular, this paper makes the following contributions:

1) We propose a NN-based emotion recognition with subnetwork nodes. The subnetwork node itself can be formed by several hidden nodes with various capabilities including feature learning, dimension reduction, etc. The subnetwork, alike neural representations in mammal cortex, can be functional as a local features extractor. The top layer of a hierarchical network, like brain, needs such subspace features produced by the subnetwork neuron to discard factors that are not relevant but, but simultaneously, recast these features into a mapping space so that the network can be performed to produce more reliable cognition. Compared with other EEG-based emotion recognition methods, the experimental results show that this subnetwork structure boosts nearly 5-10 percent accuracy of the EEG-based emotion recognition.

2) Similar to biological learning, our hierarchical learning method could use any type of features and provide a parallel and unified learning mode for multimodal psychophysiological signals. Experimental results show that our method, with multimodal signals, could provide about 91.3% accuracy, which are superior to the state-of-the-art approaches.

3) Effect of 12 channel DE features. Previous studies [2][19][18] indicate that 12 channel DE features may obtain a promising result on EEG-based emotion recognition. The experimental results of this paper are consistent with the conclusion. Furthermore, we found the DE features of eye movement also provide a better performance than other features.

II. Preliminaries and Subnetwork Nodes

A. Notations

All the notations are defined in Table I.

TABLE I
NOTATIONS TO BE USED IN THE PROPOSED METHOD

Notation	Meaning
$\mathbf{R}$	$\mathbf{R}$ represent the sets of real numbers.
M	number of training samples.
$\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M}$	$\mathbf{x}$ represents the input data and $\mathbf{y}$ represents the desired output data.
$\mathbf{a}_i$	$\mathbf{a}_i$ is the weight connecting the ith hidden nodes and the input nodes.
b_i	b_i is the bias of the ith hidden nodes.
β_i	β_i is the output weight between the ith hidden node and the output nodes.
$sum(\mathbf{e})$	$sum(\mathbf{e})$ denotes the sum of all elements of the matrix residual error $\mathbf{e}$.
$\hat{\mathbf{a}}_f^i$	input weight of the ith subnetwork node in entrance layer. $\hat{\mathbf{a}}_f^j \in \mathbf{R}^{d\times n}$
$\hat{\mathbf{a}}_h^i$	input weight of the ith subnetwork node in exit layer. $\hat{\mathbf{a}}_f^j \in \mathbf{R}^{d\times n}$
$\hat{b_f^i}$	bias of the ith subnetwork node in entrance layer $\hat{b_f^j} \in \mathbf{R}$.
$(\mathbf{a}_{fi}^j, b_f^j)$	the ith hidden node in the jth subnetwork node.
u_j	normalized function, u_j^{-1} represent its reverse function.
$\mathbf{H}_f^j$	feature data generated by j subnetwork nodes.
n	input data dimension.
m	output data dimension.
d	feature data dimension
$\mathbf{e}_L$	the residual error of current network (L subnetwork nodes).
L	the numbers of subnetwork nodes
g	g is a sigmoid or sine activation function.

B. Subnetwork nodes

Accumulated direct biological evidence supports the theory that neuron activity in the mammal's prefrontal cortex is highly heterogeneous, and the combined features extracted from mixed selectivity neurons may be central to complex behavior and cognition. Motivated by this biological evidence and the recent research developments [42][43][44], we believe that a hidden node itself can be a subnetwork formed by several nodes. In this sense, a single mapping layer can contain multiple networks. In [45], we have prove that a single-layer feedforward network with subnetwork nodes are universal approximators, especially when all the parameters of the networks are adjusted based on invertible activation functions. For M arbitrary distinct samples $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} \in \mathbf{R}^{n\times M}$ and $\mathbf{y} \in \mathbf{R}^{m\times M}$. The outputs of an SLFNs is

$$f_n(\mathbf{x}) = \sum_{i=1}^{L} \beta_i g(\mathbf{x}, \boldsymbol{a}_i, b_i) = \sum_{i=1}^{L} \beta_i \cdot \mathbf{H} \tag{1}$$

If g is invertible function, by the replacement of subnetwork nodes into the SLFNs, the mathematically model of SLFNs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

with subnetwork nodes is [45]:

$$\begin{aligned} f_L(\mathbf{x}) &= \sum_{i=1}^{L} \boldsymbol{\beta}_i u^{-1} \cdot \mathbf{H}_f^i \\ &= \sum_{i=1}^{L} \boldsymbol{\beta}_i u^{-1}(g(\hat{\mathbf{a}}_f^i \cdot \mathbf{x}_j + \hat{b}_f^i)), \hat{\mathbf{a}}_f^i \in \mathbf{R}^{n\times m}, \hat{b}_f^i \in \mathbf{R} \\ &= \sum_{i=1}^{L} \boldsymbol{\beta}_i u^{-1}(g([\mathbf{a}_{f1}^i, \cdots, \mathbf{a}_{fd}^i] \cdot \mathbf{x}_j + \hat{b}_f^i)), \mathbf{a}_{f1}^i, \cdots, \mathbf{a}_{fd}^i \in \mathbf{R}^n \end{aligned} \tag{2}$$

As seen from equation (1)-(2), we found that a subnetwork node $\hat{\mathbf{a}}_f^i$, which can be formed by several hidden nodes $[\mathbf{a}_{f1}^i, \cdots, \mathbf{a}_{fd}^i]$, could be functional as a hidden layer in a standard SLFNs. In a standard SLFNs, the dimensionality of $\mathbf{H}$ equals the number of hidden node L. But in Fig.1 (b), the dimensionality of feature data $\mathbf{H}_f^i$ follow the dimension of a subnetwork node, i.e., $\mathbf{H}_f^i \in \mathbf{R}^{d\times n}$. Fig.1 shows the architecture of the network.

III. Hierarchical network with subnetwork nodes for emotion recognition

As mentioned before, EEG signals have low signal-to-noise ratio, and are often mixed with much noise when collected. The more challenging problem is that, unlike image or speech signals, EEG signals are temporal asymmetry and nonstationary. Different from other single-classifier-based identification methods, here we study a more complex learning system for EEG signals analysis. The proposed method is composed of two parts: 1) Local features extracted from mid-level layers, 2) Feature level fusion and classification. The following subsection elaborates the architecture and its learning stages. First, a two-layer network with subnetwork nodes is carried out to extract the local features from the input data. Then these extracted features are fused together for the final classification. The structure of the proposed method is shown in Fig. 2.

Note that each hidden layer is an independent module that functions as a separated feature extractor. The proposed network structure is shown in Fig.2. Crucially, accumulated biological evidence indicates that neuron activity in the cortex is highly heterogeneous and disordered, and that the combined features extracted from mixed selectivity neurons may be central to complex behavior and cognition. Motivated by this biological evidence, we believe the following. First, an artificial neuron, which we shall call subnetwork node[38], [44], itself can be formed by several hidden nodes. Each subnetwork neuron is able to increase or decrease the dimensionality from the input data independently. Second, the outputs of each neuron, like neural representations in the mammal cortex, should be partial (not fully) connected with other neurons. Third, the outputs from each subnetwork node can be considered as specific subspace features. Useful features can be produced by recombining these subspace features with different distributions. In detail, there are several differences between our method and other multi-layer network feature selection methods:

(1) Unlike current multilayer network architectures, Fig. 2(a) shows that a subnetwork hidden node $\mathbf{a}_f^1, \mathbf{a}_n^1$ itself can be formed by hundreds of hidden nodes ($\mathbf{a}_f^1 = (a_{f1}, \cdots, a_{fd})$). Based on this architecture, the outputs of each subnetwork can be considered as subspace features. Furthermore, some multilayer methods [46], [47] require subnetwork nodes in the entrance feature layer but do not need them in the output layer in order to let all the hidden nodes fully connect. But we think that this unnaturally asymmetric architecture actually limits the learning capability. Thus, in the proposed method, subnetwork nodes are entirely instead of traditional hidden nodes.

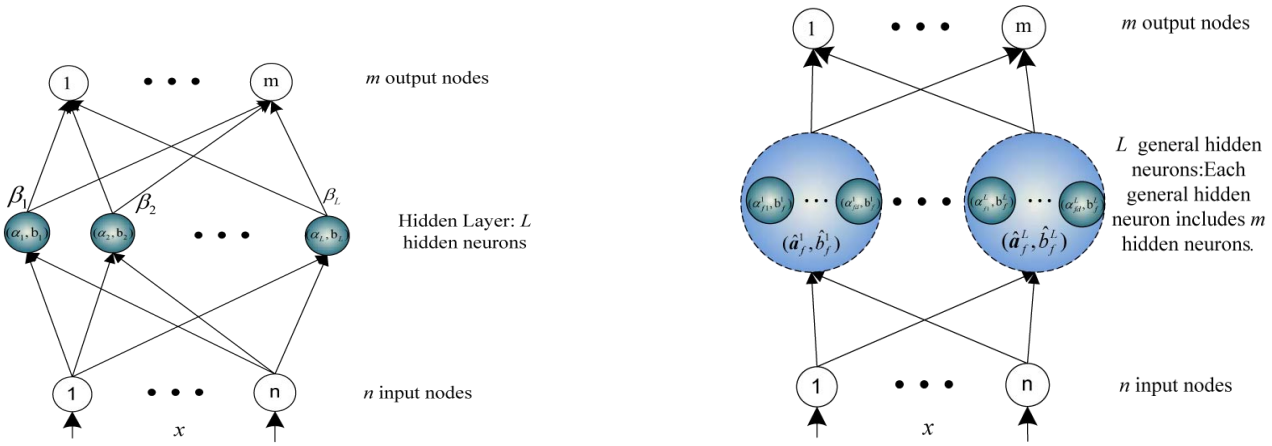
(2) Different from the current network connection principle, which states that all the hidden nodes should be fully connected (see Fig.1(b)), in our proposed architecture, each subnetwork node is only connected with its "tightly following" subnetwork. For example, in Fig.2(a), subnetwork node $\mathbf{a}_f^1$ is only connected with $\mathbf{a}_n^1$. In other words, subnetwork nodes with different subnetwork index c cannot be connected together, i.e., $\mathbf{a}_f^i$ and $\mathbf{a}_n^j$ cannot be connected together when $i \neq j$.

(3) Accumulated biological evidences show that "high-dimensional representations of a neuron with mixed selectivity allow a simple linear readout to generate a huge number of potential responses. In contrast, neural representations based on highly specialized neurons are low-dimensional". This evidence is highly consistent with the domain assumption in machine learning area that useful feature data intrinsically exists in several subspaces. Unlike current multilayer/auto-encoder methods in which features extracted from the entire mid-layer, we belief the neural representations (outputs from each subnetwork node) should be mixed with diverse distributions/manners based on the above biological evidences. In Fig.2(b)p, we show that how are the subspace features extracted and combined.
(4) Unlike other hierarchical networks which include hundreds of layer to generate deep features, the generic features are obtained from two general layers, which greatly reduce the network depth and computational workloads. It should be note that there are several million parameters in the first general layers, which is not a small network. As seen in Fig.2, the first general layers include several two-layer networks (Part I). And each subnetwork node (Fig.1(b)) in the two-layer network includes hundreds of hidden nodes.
(5) The iterative methods used in DL suffer from converging slowly, getting trapped in a local minimum, and being sensitive to the learning rate setting. Unlike BP-based iterative methods, in this paper, the Moore-Penrose generalized inverse is used for parameter calculation. By doing so, each subnetwork node in the system does not need to retrain iteratively (see Step 1-7), which also boost the learning speed.

A. Data preprocessing

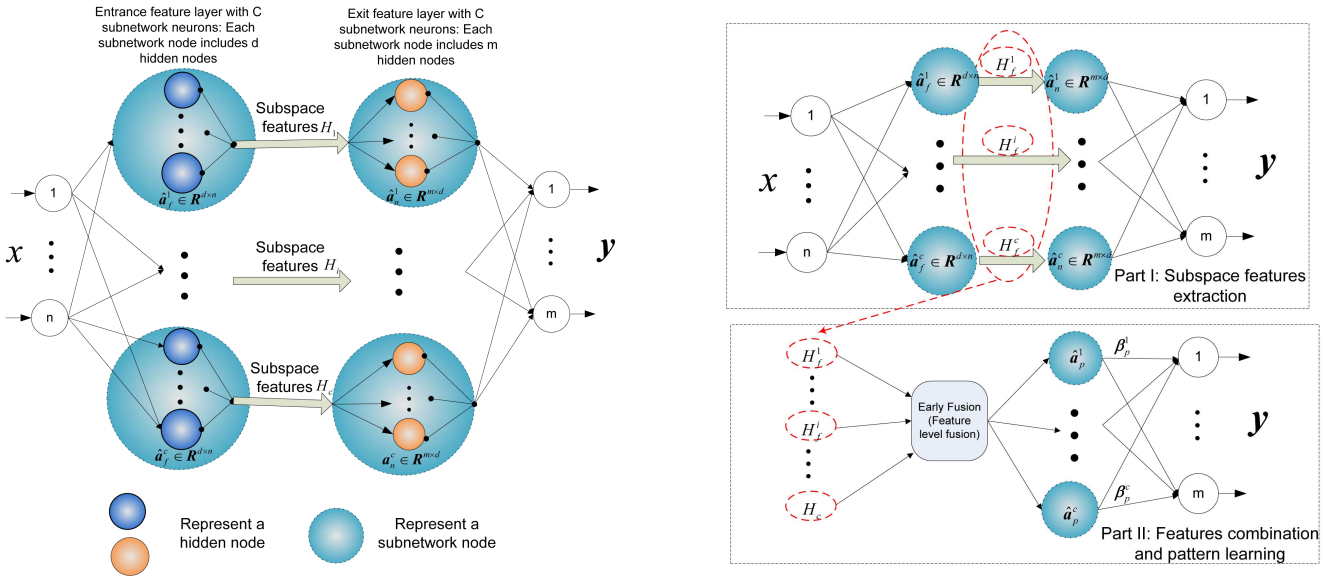
According to the feedback of the subjects, only the experiments when the targeted emotions were evoked were selected for further examination. Similar to [18], the raw

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems



(a) Standard SLFN structure (b) SLFN with subnetwork nodes

Fig. 1. Difference and relationship among standard SLFN and our sturcture.



(a) Two-layer network with subnetwork nodes (b) our EEG-based learning framework

Fig. 2. Difference and relationship among a standard two-layer network and our method.

EEG data signals were visually checked by removing EMG and EOG signals manually. To filter noises and artifacts, the EEG signals are dealt with a bandpass filter between 0.3 to 50 Hz. After this, an EEG segment is extracted from the duration of each movie correspondingly. Each channel data (totally 62 channels) is then divided into the same-length epochs of 1 second.

According to the previous studies, DE has a promising capability of recognition EEG patterns between low and high frequency energy [19]. The DE calculation formula is:

$$h(X) = -\int_X f(x)\log(f(x))dx \tag{3}$$

If the time series X obeys the Gauss distribution $N(\mu,\delta)$, the DE features can be obtained by:

$$\begin{aligned} h(X) &= -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{(2\pi\delta^2)}} e^{-\frac{(x-\mu)^2}{2\delta^2}} \log\left(\frac{1}{\sqrt{(2\pi\delta^2)}} e^{-\frac{(x-\mu)^2}{2\delta^2}}\right) dx \\ &= \frac{1}{2}\log(2\pi e\delta^2) \end{aligned} \tag{4}$$

According to [18], DE features can be obtained in five frequency bands (1-3 Hz, 4-7 Hz, 8-13 Hz, 14-30 Hz, 31-50 Hz).

B. Local features extraction with subnetwork nodes

In this subsection, we train the two-layer network architecture shown in Fig.2(a) and Fig.2(b)(Part I) to obtain subspace local features.

Step 1: Given M arbitrary distinct training samples $\{(\mathbf{x}_k,\mathbf{y}_k\}_{k=1}^{M}, \mathbf{x}_k \in \mathbf{R}^n$ are sampled from a continuous system.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

The initial subnetwork node of the entrance layer are obtained using orthogonal random:

$$\mathbf{H}_f^c = g(\hat{\mathbf{a}}_f^c, \hat{b}_f^c, \mathbf{x}),\ (\hat{\mathbf{a}}_f^c)^T \cdot \hat{\mathbf{a}}_f^c = \mathbf{I},\ (\hat{b}_f^c)^T \cdot \hat{b}_f^c = 1 \tag{5}$$

where $\hat{\mathbf{a}}_f \in \mathbf{R}^{d\times n}$, $\hat{b}_f \in \mathbf{R}$ is the orthogonal random weight and bias of the entrance mapping layer. $\mathbf{H}_f^c$ is the c-th subspace features. c represents subnetwork node index and initial index $c = 1$.

Step 2: Given an invertible activation function g, obtain the subnetwork node of the exit feature layer $(\hat{\mathbf{a}}_h^c, \hat{b}_h^c)$ by

$$\begin{aligned} \hat{\mathbf{a}}_h^c &= g^{-1}(u_n(\mathbf{y})) \cdot (\mathbf{H}_f^c)^{-1}\ , \hat{\mathbf{a}}_h^c \in \mathbf{R}^{d\times m} \\ \hat{b}_h^c &= \sqrt{mse(\hat{\mathbf{a}}_h \cdot \mathbf{H}_f^c - g^{-1}(u_n(\mathbf{y})))}\ , \hat{b}_n^c \in \mathbf{R} \end{aligned} \tag{6}$$

where $\mathbf{H}^{-1} = \mathbf{H}^T(\frac{C1}{\mathbf{I}} + \mathbf{H}\mathbf{H}^T)^{-1}$; $C1 > 0$ is a regularization value; u_n is a normalized function $u_n(\mathbf{y}) : \mathbf{R} \to (0, 1]$; g^{-1} and u_n^{-1} represent their reverse function.

Step 3: Update the output error $\mathbf{e}_c$ as

$$\mathbf{e}_c = \mathbf{y} - u_n^{-1} g(\mathbf{H}_f^c, \hat{\mathbf{a}}_h^c, \hat{b}_h^c) \tag{7}$$

We can get error feedback data $\mathbf{P}_c = g^{-1}(u_n(\mathbf{e}_c)) \cdot (\hat{\mathbf{a}}_h^c)^{-1}$.

Step 4: Update the subnetwork node $\hat{\mathbf{a}}_f^c, \hat{b}_f^c$ in the entrance layer

$$\begin{aligned} \hat{\mathbf{a}}_f^c &= g^{-1}(u_j(\mathbf{P}_{c-1})) \cdot \mathbf{x}^{-1}\ , \hat{\mathbf{a}}_f^c \in \mathbf{R}^{n\times d} \\ \hat{b}_f^c &= \sqrt{mse(\hat{\mathbf{a}}_f^c \cdot \mathbf{x} - \mathbf{P}_{c-1})}\ , \hat{b}_f^c \in \mathbf{R} \end{aligned} \tag{8}$$

Step 5: obtain the c-th subspace feature data

$$\mathbf{H}_f^c = g(\mathbf{x}, \hat{\mathbf{a}}_f^c, \hat{b}_f^c) \tag{9}$$

Step 6: Set $c = c + 1$, add a new *subnetwork node* $\hat{\mathbf{a}}_f^c, \hat{b}_f^c$ in the feature mapping layer with orthogonal random initialization (equation (5)).

Step 7: Repeat steps 2 to 6 $L-1$ times, then obtain the L subspace features $\{\mathbf{H}_f^1, \cdots, \mathbf{H}_f^L\}$.

C. Features fusion

To make synergistic use of the emotion recognition, features extract from multiple modality (e.g., eye, EEG, skin, etc.) are combined through early fusion. literature [48] [49] indicate that if the data contain corrected information, early fusion is beneficial over later fusion by a simple union of different features into one super-vector. For example, there are two sets of subspace features which have been extracted from two different networks. Here, we redefine the features coming from network #1 as $\mathbf{H}^1 = \{\mathbf{H}_1^1, \mathbf{H}_2^1, \cdots, \mathbf{H}_c^1\}$, and those from network #2 as $\mathbf{H}^2 = \{\mathbf{H}_1^2, \mathbf{H}_2^2, \cdots, \mathbf{H}_c^2\}$. The combination features can be obtained by early fusion:

$$\mathbf{H}^{1\oplus 2} = [\mathbf{H}_1^1,\ \mathbf{H}_2^1,\ \cdots,\ \mathbf{H}_c^1,\ \mathbf{H}_1^2,\ \mathbf{H}_2^2,\ \cdots,\ \mathbf{H}_c^2]^T \tag{10}$$

Furthermore, motivated by anonymous reviewer, we introduce maxpooling into the proposed method. The past few years have witnessed the bloom of Convolutional Neural Network [26], [27], [28]. In many well-known CNN models like GoogLeNet [50], AlexNet [51], etc., maxpooling is widely used for feature combination and dimension reduction. Inspired by anonymous reviewer, here we introduce maxpooling into our proposed method for feature fusion.

$$\mathbf{H}^{1\oplus 2} = max(\mathbf{H}^1, \mathbf{H}^2) \tag{11}$$

Furthermore, motivated by anonymous reviewer, we introduce maxpooling into the proposed method. The past few years have witnessed the bloom of Convolutional Neural Network [26], [27], [28]. In many well-known CNN models like GoogLeNet [50], AlexNet [51], etc., maxpooling is widely used for feature combination and dimension reduction. Inspired by anonymous reviewer, here we introduce maxpooling into our proposed method for feature fusion.

$$\mathbf{H}^{1\oplus 2} = max(\mathbf{H}^1, \mathbf{H}^2) \tag{12}$$

Given several features $\mathbf{H}^1, \cdots, \mathbf{H}^c$, K represents a combination operator, the combined features can be expressed as

$$\begin{aligned} \mathbf{H}^{1\oplus 2} &= K(\mathbf{H}^1, \mathbf{H}^2) \\ \mathbf{H}^{1\oplus 2\oplus 3} &= K(K(\mathbf{H}^1, \mathbf{H}^2), \mathbf{H}^3) \\ &\vdots \\ \mathbf{H}^{1\oplus 2\oplus \cdots \oplus c} &= K(\cdots K(K(\mathbf{H}^1, \mathbf{H}^2), \mathbf{H}^3) \cdots) \end{aligned} \tag{13}$$

Fig.3 indicates the framework map representations from input EEG data to c low-dimensional subspace features, and to a high-level image combined features, used for categorization. The image representation begins with EEG features from which local descriptors, such as DE or other EEG features, are extracted to create a powerful representation. Current accumulated biological evidence [43] shows that the investigations of mixed neurons have started to point out their importance, both for the implementation of brain functions and for coding. The brain needs subspace features produced by a neuron to remove relevant factors, but, meanwhile, to recast the subspace features into a mapping space in order to generate complex and stable behavior. Fig. 3 shows that the learning structure and dimensionality correspond with to the major principals of the biological evidence mentioned previously. In the hierarchical architecture, the subspace feature dimensionality extracted from a neuron decreases progressively. At the combination level, the training samples are put through an early fusion method for a final classifier. The graph (Fig. 3) illustrates the trends of the dimensionality of the representation through the various processes in the framework.

Powered by our subnetwork nodes, any type of features can be directly extracted and combined. Our method, with its multiple features, can be summarized in Fig.5 in the following subsection.

D. Classification for emotion recognition

The feature extraction and fusion steps descried above contain optimized features which wait for classification. In this subsection, we focus on classifying the subspace features extracted from the mid-layer NN. As shown in Fig.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

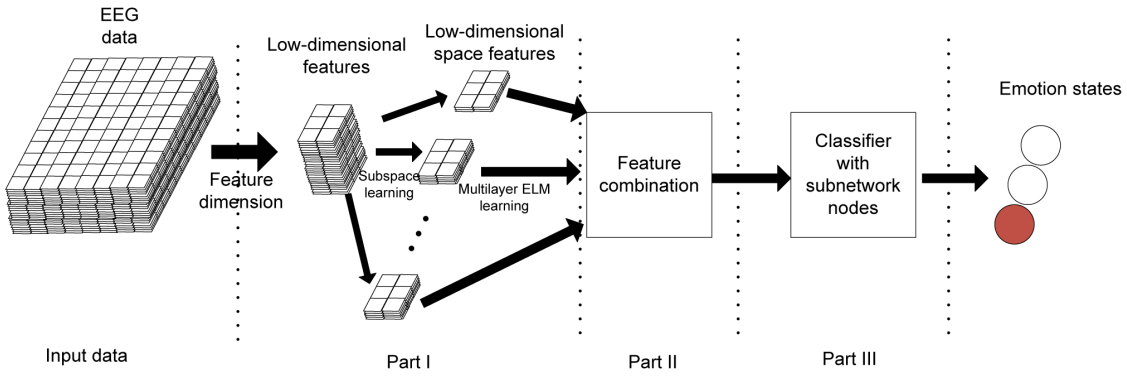


Fig. 3. The proposed learning system from EEG data to low-dimensional features, subspace low-dimensional features, and to a mid-level dimensional features, which is used for emotion recognition.

2(b)(part II), a classifier with subnetwork nodes [45] is used for the final classification.

Theorem 1: [45] Given N arbitrary distinct samples $\{(\mathbf{x}_i, \mathbf{t}_i)_{i=1}^N\}$, $\mathbf{x}_i \in \mathbf{R}^n, \mathbf{t}_i \in \mathbf{R}^m$, a sigmoid or sine activation function g, and then for any continuous desired outputs $\mathbf{t}$, we have $\lim_{c\to+\infty} \|\mathbf{t} - (u^{-1}(g(\hat{\mathbf{a}}_p^1 \cdot \mathbf{x} + \hat{b}_p^1)) \cdot \boldsymbol{\beta}_p^1 + \cdots + u^{-1}(g(\hat{\mathbf{a}}_p^c \cdot \mathbf{x} + \hat{b}_p^c)) \cdot \boldsymbol{\beta}_p^c)\| = 0$ holds with probability one if

$$\begin{aligned} &\hat{\mathbf{a}}_p^c = g^{-1}(u(\mathbf{e}_{n-1})) \cdot \mathbf{x}^T (\frac{C2}{\mathbf{I}} + \mathbf{x}\mathbf{x}^T)^{-1}, \hat{\mathbf{a}}_p^c \in \mathbf{R}^{n\times m} \\ &\hat{b}_p^c = sum(\hat{\mathbf{a}}_n^c \cdot \mathbf{x} - g^{-1}(u(\mathbf{e}_{n-1})))/N, \hat{b}_p^c \in \mathbf{R} \\ &g^{-1}(\cdot) \begin{cases} arcsin(\cdot) & \text{if } g(\cdot) = sin(\cdot) \\ -\log(\frac{1}{(\cdot)} - 1) & \text{if } g(\cdot) = 1/(1+e^{-(\cdot)}) \end{cases} \end{aligned} \tag{14}$$

$$\boldsymbol{\beta}_p^c = \frac{\langle \mathbf{e}_{n-1}, u^{-1}(h(\hat{\mathbf{a}}_n^c \cdot \mathbf{x} + \hat{b}_n^c))\rangle}{\|u^{-1}(h(\hat{\mathbf{a}}_n^c \cdot \mathbf{x} + \hat{b}_n^c))\|^2} \tag{15}$$

where $\mathbf{x}^T(\frac{C2}{\mathbf{I}} + \mathbf{x}\mathbf{x}^T)^{-1} = \mathbf{x}^{-1}$ is the Moore-Penrose generalization inverse of the training samples; g^{-1} represents its inverse function; u is a normalized function $u(x): \mathbf{R} \to (0,1]$ which processes input x and target data by mapping it from its original range to the range $(0,1]$; u^{-1} is a inverse function of u, which processes target data and input x by mapping it from the range $(0,1]$ to its original range. [1]

Furthermore, other classifiers, such as SVM, can be used in the method as well. The proposed algorithm could be summarized in the following Algorithm 1-2.

Algorithm 1 The proposed method for single modality

Given a large training dataset $\{(\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^M, \mathbf{x}_k \in \mathbf{R}^n$, an invertible activation function g, number of hidden nodes in each subnetwork node d (d equals number of targeted dimensionality of the subspace features), regularization coefficient C, and the number of subnetwork nodes L:

Part I: Subspace feature extraction:

Step 1: Set $c = 1$, randomly generate the subnetwork node for entrance feature layer by equation (5).

while $c < L$ **do**

Step 2: Calculate the subnetwork node for exit feature layer by equation (6)

Step 3: Calculate the output error and error feedback data by equation (7)

Step 4: Update the subnetwork node $\hat{\mathbf{a}}_f^c, \hat{b}_f^c$ in the entrance layer by equation (8)

Step 5: obtain the c-th subspace feature data by equation (9)

Step 6: Set $c = c+1$, add a new *subnetwork node* $\hat{\mathbf{a}}_f^c, \hat{b}_f^c$ in the feature mapping layer with orthogonal random initialization (equation (5)).

Step 7: Repeat steps 2 to 6 $L-1$ times, obtain the L subspace features $\{\mathbf{H}_f^1, \cdots, \mathbf{H}_f^L\}$.

end while

Obtain c subspace features $\mathbf{H} = \{\mathbf{H}_f^1, \cdots, \mathbf{H}_f^L\}$.

Part II: Pattern learning: Given fusion feature $\mathbf{H}$ and corresponding desire output $\mathbf{t}$, set $c = 1, e_1 = \mathbf{t}$.

while $c < L$ **do**

Step 1: Calculate the cth subnetwork hidden node $(\hat{\mathbf{a}}_p^c, \hat{b}_p^c)$, and output weights $\boldsymbol{\beta}_p^c$ by equation (14)-(15)

Step 2: Calculate $\mathbf{e}_c = \mathbf{e}_{c-1} - \boldsymbol{\beta}_p^c \cdot u^{-1} g(\hat{\mathbf{a}}_p^c, \hat{b}_p^c, \mathbf{x})$.

end while

IV. Performance Verification

In this section, we test our method on two different EEG datasets[2]. The experiments are conducted in Matlab 2014 with 32 GB of memory. In the following subsection, we conduct comparative experiments of our method with six methods for EEG-based emotion recognition. The six classification methods are as follows:

1) DBN [32]
2) Extreme Learning Machine (ELM) [52]
3) SVM [53]
4) Hierarchical ELM (H-ELM) [47]
5) KNN
6) Linear Regression (LR)

Furthermore, in order to compare the performance for multi-source fusion, some fusion methods are set as the rival methods:

[1] In Matlab environment, we can use MATLAB commend mapminmax to utilize u and u^{-1}.

[2] http://bcmi.sjtu.edu.cn/~seed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

Algorithm 2 The proposed method for multiple modality

Given N single features groups $(Q_1, \cdots, Q_N)$ extracted from the same dataset $Q_1 = \{(\mathbf{x}_k^1, \mathbf{y}_k^1)\}_{k=1}^M, \mathbf{x}_k^1 \in \mathbf{R}^{n_1}, \cdots, Q_N = \{(\mathbf{x}_k^N, \mathbf{y}_k^N)\}_{k=1}^M, \mathbf{x}_k^N \in \mathbf{R}^{n_N}$ (the dimensionality of each features group do not need to be equal, which means $n_1, \cdots, n_N$ do not need to be equal), an invertible activation function g, number of hidden nodes in each subnetwork node d, regularization coefficient C, and the number of subnetwork nodes c. Set $c = 1$.

Layer 1: Subspace features extraction

for $c < N$ **do**

Obtain the L subspace features based on Algorithm 1.Part I by using group data Q_c.

end for

obtain $N \times L$ subspace features $\{\mathbf{H}_f^1, \cdots, \mathbf{H}_f^{N \times L}\}$.

Layer 2: Subspace features combination

Obtain combination features $\mathbf{H}$ as:

$$\mathbf{H} = \mathbf{H}^{1 \oplus 2 \oplus \cdots \oplus (N)} \tag{16}$$

Layer 3: Obtain simulated outputs based on Algorithm 1 Part II.

1) Decision level fusion: maximal rule and sum rule [2]
2) Feature level fusion: fuzzy integral fusion [54]

The codes used for DBN, SAE, LLP, and Hierarchical ELM are downloaded from the Internet. The parameters in any learning method can be tuned for each experiment.

A. Data processing and experimental environment setting

Previous studies [19][40] have already tested the reliability of film clips (see Fig.6(a)) to elicit emotions. In our work, we use the same datasets released by [18][2]. There are in total fifteen clips in one experiment, and each of them lasts for about 4 min. There are three categories of emotion (Positive, Neutral, and Negative) evaluated, where each emotion has five corresponding emotional clips.

The first EEG dataset (SEED) is released by [18]. Fourteen subjects (7 males and 7 females), with self-reported normal or corrected-to-normal vision and normal hearing, participated in the experiments. Fig.6 shows the experiment scene. Each subject participated in the experiment three times at an interval of one week or longer. There is total of three sessions (3×14 experiments) evaluated here.

To further compare the generalization performance, we compute differential asymmetry (DASM) and rational asymmetry (RASM) features [18] as differences and ratios between the DE features. DASM, RASM, and DCAU features are, respectively, defined as follows:

$$DASM = DE(\mathbf{X}_{left}) - DE(\mathbf{X}_{right}) \tag{17}$$

$$RASM = DE(\mathbf{X}_{left}) / DE(\mathbf{X}_{right}) \tag{18}$$

$$DCAU = DE(\mathbf{X}_{frontal}) / DE(\mathbf{X}_{posterior}) \tag{19}$$

where $\mathbf{X}_{left}$ and $\mathbf{X}_{right}$ represent the pairs of electrodes on the left and right hemisphere. $\mathbf{X}_{frontal}$ and $\mathbf{X}_{posterior}$ represent the pairs of frontal-posterior electrodes. The detailed information about SEED dataset are shown in the following Table II.

Different from the first EEG dataset, the second one has EEG data with eye movement information [2]. Fifteen video clips, same in SEED dataset, are used for each experiment. Nine healthy, right-handed subjects (5 females and 4 males) participated in the experiment. Each of them took part in the experiment three times at an interval of about one week, and there is a total of 27 experiments evaluated here. All the subjects are undergraduate or graduate students, aged between 20 and 24 years, with normal or corrected-to-normal vision, and none of them have any history of mental disease or drug use. Eye movement signals are recorded using SMI ETG eye tracking glasses. EEG signals are recorded with a 1000 Hz sampling rate using ESI NeuroScan System. In the experiment, we use DE eye movement features which are shown in the following Table III.

In this section, we systematically estimate the generalization performance of six classifiers, logistic regression (LR), K nearest neighbor (kNN), support vector machine (SVM), extreme learning machine (ELM), hierarchical ELM (H-ELM) deep belief networks(DBNs), and the proposed method. These classifiers utilize the five aforementioned features as inputs. Similar to [18], we use the same range of parameters: For kNN, we use $k = 5$ for baseline. For LR, we use L2-regularized LR and adjust the regularization parameter in $[1.5 : 0.5 : 10]$. For SVM and ELM, optimal parameters are selected from the space $[2^{-10}, 2^{-9}, \cdots, 2^{10}]$ in each experiment. For H-ELM, 300, 300, and 1000 hidden neurons (N1=N2=300,N3=1000) are used in the first, second and third layer. For DBN, we use two hidden layers with epoch 1000, the parameter batch size equals 201, the parameter momentum, unsupervised, and supervised learning rate equals to 0.1, 0.5, and 0.6, respectively. For each experiment, the optimal number of neurons at the first and the second layer of DBN is selected from the ranges of $[200 : 500]$ and $[150 : 500]$, respectively. For our method, regularization parameter $C2$ is selected from the space $[2^{-10}, 2^{-9}, \cdots, 2^{10}]$ for each experiment, in order to be consistent with ELM/SVM, while parameter $C1$ is selected from the same space $[2^{-10}, 2^{-9}, \cdots, 2^{10}]$ for each session (i.e., value of $C1$ should be the same among all fifteen experiments). The fusion strategy in our method include early fusion and maxpooling. Table IV shows the detailed experimental setting.

B. Subject dependent test

Subject dependent is to predict the same person's emotion based on his/her previous responses from different stimulus. The training and the testing samples come from different sessions of the same experiment. In this experiment, the training samples contains the first nine sessions, while the test data includes the later six sessions (totally 15 sessions). In order to be consistent with the previous

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

No.	Emotion label	Film clips sources
1	negative	Tangshan Earthquake
2	negative	1942
3	positive	Lost in Thailand
4	positive	Flirting Scholar
5	positive	Just Another Pandora's Box
6	neutral	World Heritage In China

(a) Details of film clips used in the experiment

(b) the experiment scene

Fig. 4. Details of the experiment

TABLE II
DETAILED INFORMATION OF SEED DATASET

Features	Clip#1	Clip#2	Clip#3	Clip#4	Clip#5	Clip#6	Clip#7	Clip#8	Clip#9	Clip#10	Clip#11	Clip#12	Clip#13	Clip#14	Clip#15
ASM	235×270	233×270	206×270	238×270	185×270	195×270	237×270	216×270	265×270	237×270	235×270	233×270	235×270	238×270	206×270
DASM	235×135	233×135	206×135	238×135	185×135	195×135	237×135	216×135	265×135	237×135	235×135	233×135	235×135	238×135	206×135
DCAU	235×115	233×115	206×115	238×115	185×115	195×115	237×115	216×115	265×115	237×115	235×115	233×115	235×115	238×115	206×115
DE	235×310	233×310	206×310	238×310	185×310	195×310	237×310	216×310	265×310	237×310	235×310	233×310	235×310	238×310	206×310
PSD	235×310	233×310	206×310	238×310	185×310	195×310	237×310	216×310	265×310	237×310	235×310	233×310	235×310	238×310	206×310
RASM	235×135	233×135	206×135	238×135	185×135	195×135	237×135	216×135	265×135	237×135	235×135	233×135	235×135	238×135	206×135

1. Clip#i represents the ith Clip.
2. We use the term of $X \times Y$ to define sample numbers and dimensions. For example, 235×270 represents this feature group has 235 samples with 270 dimensions.

TABLE III
DETAILED INFORMATION OF EYE MOVEMENT DATA

Features	Clip#1	Clip#2	Clip#3	Clip#4	Clip#5	Clip#6	Clip#7	Clip#8	Clip#9	Clip#10	Clip#11	Clip#12	Clip#13	Clip#14	Clip#15
Eye blink	58×4	58×4	51×4	59×4	46×4	48×4	59×4	54×4	66×4	59×4	58×4	58×4	58×4	59×4	51×4
Eye saccade	58×8	58×8	51×8	59×8	46×8	48×8	59×8	54×8	66×8	59×8	58×8	58×8	58×8	59×8	51×8
Fixation	58×4	58×4	51×4	59×4	46×4	48×4	59×4	54×4	66×4	59×4	58×4	58×4	58×4	59×4	51×4
Pupil diameter	58×8	58×8	51×8	59×8	46×8	48×8	59×8	54×8	66×8	59×8	58×8	58×8	58×8	59×8	51×8
Pupil dispersion	58×8	58×8	51×8	59×8	46×8	48×8	59×8	54×8	66×8	59×8	58×8	58×8	58×8	59×8	51×8

1. Clip#i represents the ith Clip.
2. We use the term of $X \times Y$ to define sample numbers and dimensions. For example, 58×4 represents this feature group has 58 samples with 4 dimensions.

TABLE IV
NETWORK CONFIGURATION

Methods	parameter details
SVM	Linear Kernel, search space $2^{[-10:10]}$ with a step of one.
KNN	Baseline k equals 5
ELM	1000 hidden neurons, search space $2^{[-10:10]}$ with a step of one.
H-ELM	N1=N2=300, N3=1000, search space for $C1$ and $C2$ is $2^{[-10:10]}$ with a step of one.
DBN	the optimal number of neurons at the first and the second layer of DBN is selected from the ranges of [200 : 500] and [150 : 500], respectively.
OURS	search space for $C1$ and $C2$ is $2^{[-10:10]}$ with a step of one. Three subnetwork nodes. In each subnetwork node, 500 hidden nodes are used.

TABLE V
THE MEAN ACCURACY OF EEG FEATURES FROM FULL CHANNELS

Methods	DE	PSD	DASM	DCAU	RASM
SVM	83.99	59.60	72.81	77.38	74.74
DBN	86.08	61.90	72.73	77.20	71.30
KNN	72.60	-	-	-	-
LR	82.70	-	-	-	-
ELM	82.92	60.80	70.17	77.08	72.47
Ours	**93.26**	73.81	87.09	89.28	87.50

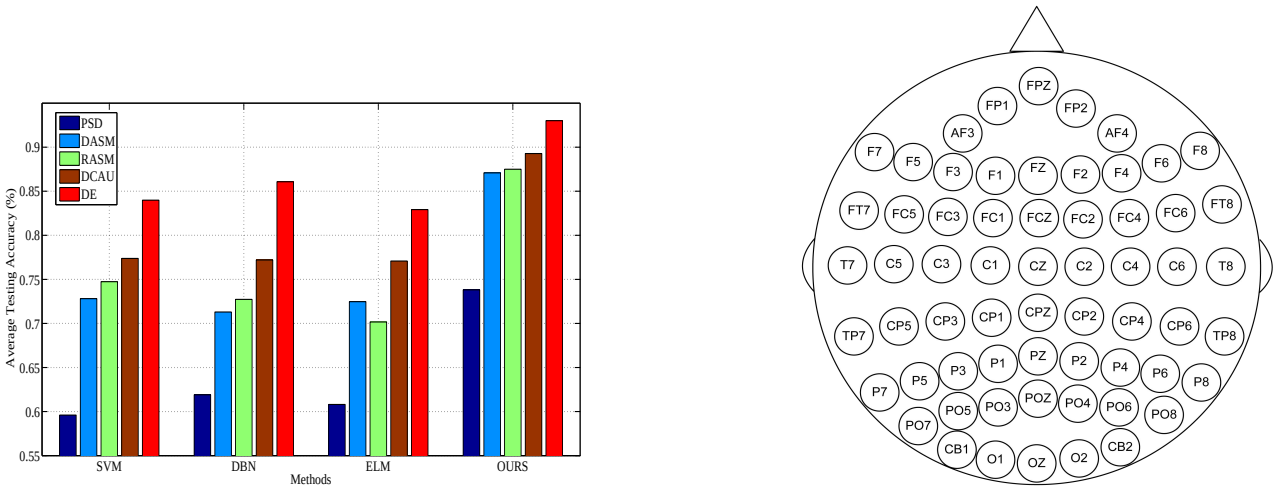
TABLE VI
THE MEAN ACCURACY OF EEG FEATURES FROM 12 CHANNELS

Methods	DE	PSD	DASM	DCAU	RASM
SVM [18]	86.65	62.92	75.86	71.82	75.70
ELM	85.09	67.11	71.81	72.84	75.01
Ours	**91.51**	83.93	87.50	83.71	86.90

studies, we only utilize the first and the third session (2×14 experiments) from the SEED dataset.

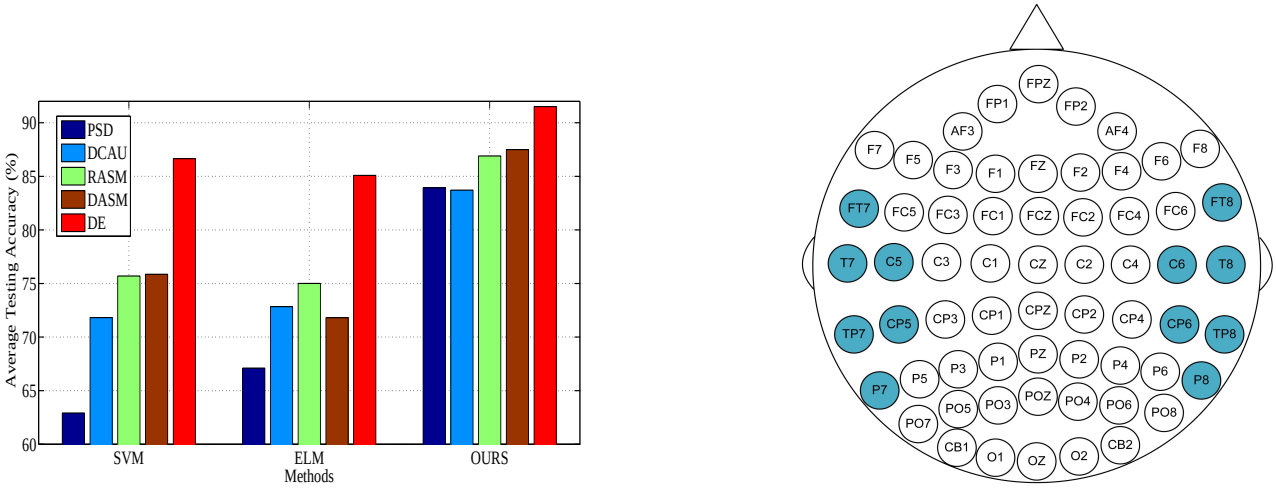
To show the profit of our method for emotion recognition performance, comparison tests have been carried out about the accuracy of the proposed method. Table V displays the recognition accuracy comparison of DBN, ELM, KNN, LR, SVM, and the proposed method. As seen from the Tables, the profit of our approach for recognition accuracy is obvious. Furthermore, Fig.5-6 shows the comparison performance by using different single features. The experimental results indicate that the proposed algorithm consistently outperforms all the compared algorithms on the EEG-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems



(a) (b)

Fig. 5. (a) Comparison experiment results with 62 selected channels. (b) Profiles of full 62 selected channels.



(a) (b)

Fig. 6. (a) Comparison experiment results with 12 selected channels. (b) Profiles of 12 selected channels: FT7, FT8, T7, T8, C5, C6, TP7, TP8, CP5, CP6, P7, and P8.

based emotion recognition. In addition, the experimental results are consistent with the previous works [18], [19], which show that the DE feature almost provide the best performance of EEG-based recognition.

Although [18] mentioned that with 12 selective channels, SVM obtains a little bit higher accuracy than that of DBN/SVM with original full 62 channels, where the remaining 50 channels are not "uninformative". These statements are consistent with our experimental results as well. As seen from Table V-VI, the performance of full 62 channels obtained by our method is approaching nearly 93%, higher than the performance of 12 channels profile (91.5%). As per our knowledge, [18] is the current state-of-the-art results on the dataset. From these figures, it can be deduced that our approach outperforms the other current leading methods.

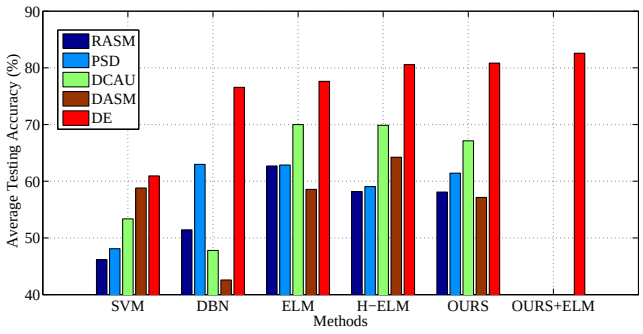


Fig. 7. Comparison experiment results of subject dependent test

C. Cross session test

Cross session is to predict the same person's emotion at a different time when the same stimuli is received, i.e.,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

stability of emotion recognition model over time. In this test, the first two sessions (two different days) from the same subjects are used as training data, and then, the remaining one session is used for test data. For H-ELM, regularization parameter $C2$ is selected from the space $[10^{-10}, 2^{-9}, \cdots, 10^{10}]$ for each experiment, while parameter S is selected from the $[0.1, 0.2, \cdots, 1]$ for a session. For ELM, SVM, DBN, and our proposed method, the way of parameters selection is the same as the subject dependent test.

TABLE VII
THE MEAN ACCURACY OF DIFFERENT KINDS OF FEATURES

Methods	DE	PSD	DASM	DCAU	RASM
full selected channels					
SVM	60.95	48.10	58.81	53.37	46.19
DBN	76.57	62.98	42.59	47.80	51.43
ELM	77.62	62.86	58.57	70.00	62.70
H-ELM	76.19	-	-	-	-
Ours	80.84	61.43	57.14	67.14	58.10
Ours+ELM	**82.86**	-	-	-	-
12 selected channels					
H-ELM	80.57	59.05	64.24	69.87	59.87
Ours	78.08	-	-	-	-
Ours+ELM	80.24	-	-	-	-

To show the profit of our method on cross session test, the results obtained by ELM, H-ELM, SVM, DBN, and Ours are showed in this subsection. Table VII shows the performance evaluation of the proposed method and other classifiers. As seen from the Table VII and Fig.7, the profit of the proposed method for testing accuracy is obvious. It should be noticed that if ELM replaces our method in the top layer, the best performance 82% will be achieved.

D. One classifier for all users test

Different from above subject dependent and cross session tests (one network per user), we try to predict emotions by one network, i.e., one trained network for all users' emotion prediction. Different from subject dependent test, in this test, all the training data from the first nine clips are used for training a network, while the test data from the later six clips are used for performance evaluation. In detail, we use $2 \times 15 \times 9$ data groups for one network training, and utilize $1 \times 15 \times 6$ data groups for testing purpose. Thus the total sample-numbers of EEG features for the training and testing data is 56280 (2010×2(times) $\times$ 14(persons)), and 19376 ($1384 \times 1 \times 14$(persons)).

Fig.8 and Table V display the performance comparison of H-ELM, ELM, and the proposed method. Based on the pervious experimental results (Fig.5-7), the generalization performance of SVM and LR are obviously weaker than ELM-based classifiers. Thus SVM and LR are not included in this challenge test. As seen from Table V and Fig.8, the experimental performance of our method consistently better than all the compared algorithms on all the types of features.

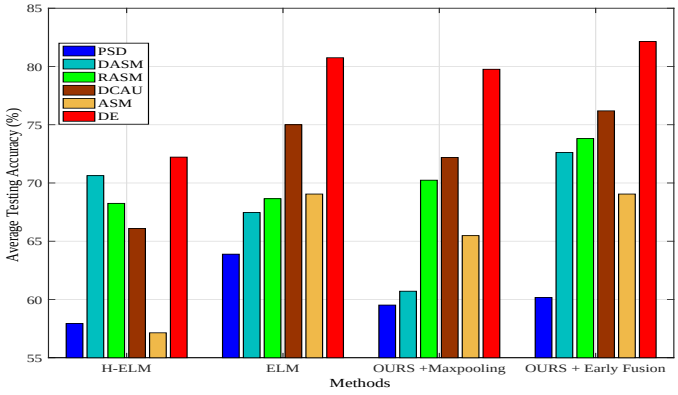


Fig. 8. Comparison experiment results of one classifier for all users test

TABLE VIII
PERFORMANCE COMPARISON OF ONE CLASSIFIER FOR ALL USERS TEST

Methods	DE	PSD	DASM	RASM	DCAU	ASM
full selected channels						
H-ELM	72.22	57.94	70.63	68.25	66.10	57.14
ELM	80.75	63.89	67.46	68.65	75.10	69.05
Ours+maxpooling	79.76	59.56	60.71	70.24	72.19	65.48
Ours+early fusion	**82.14**	60.17	72.62	73.81	76.19	69.05
12 selected channels						
ELM	82.94	-	-	-	-	
Ours	**85.71**	-	-	-	-	

TABLE IX
THE MEAN ACCURACY FOR DIFFERENT KINDS OF FEATURES (MEAN: AVERAGE TESTING ACCURACY)

Methods	DE	PSD	DASM	RASM	DCAU	ASM
full selected channels						
ELM	64.00	56.77	42.51	44.80	54.37	55.10
Ours+maxpooling	**78.79**	51.07	56.78	54.58	51.07	52.52
Ours+early fusion	74.10	52.77	65.47	58.59	57.01	54.12

E. one classifier for emotion recognition with changed times, persons, and simulations

Motivated by anonymous reviewers that the neuron activities are heterogeneous and nonstationary over time and space, here we try to highlight advantages of our method over other classifiers by a more tough test. Actually in SEED dataset, three factors including persons, measure-

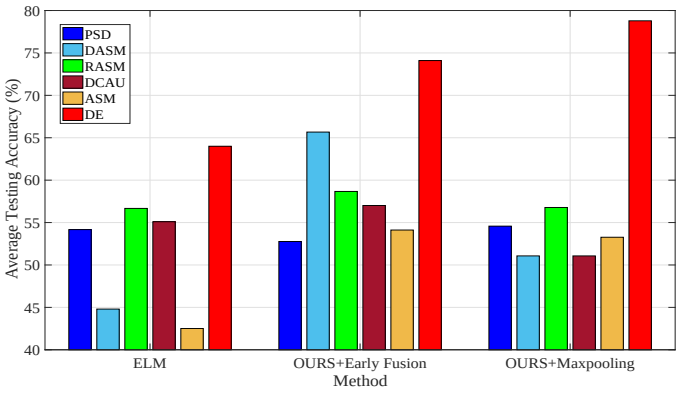
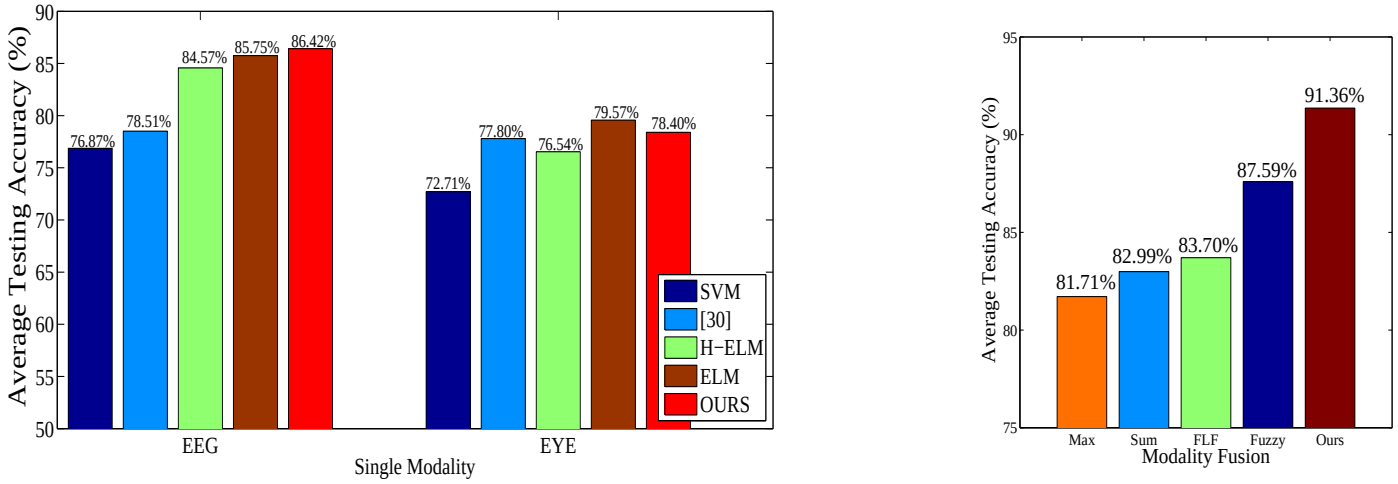


Fig. 9. Comparison experiment results (one classifier for all users with changed times, persons, and simulations

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems



(a) Performance of each single modality with different methods

(b) Comparison results with different fusion strategies

Fig. 10. Performance difference between single modality and multiple modality.

ment times, and stimulations are considered. We want to test the emotion recognition capability under the condition that all the three factors are changed. Thus this test can be considered as a combination of all the above-mentioned tests. Because we have total 14 persons, 15 clips/sessions, and 3 sessions in SEED dataset, we select training data from the first nine persons, the first nine clips/simulations, and the first two sessions, i.e., we have $9\times9\times2$ experiments. So the total sample-numbers of EEG features for the training and testing data is 36180 (2010×2(times)$\times9$(persons)), and 6920 (1384×1(times)$\times5$(persons)).

Fig.9 and Table IX display the performance comparison of ELM, and the proposed method. Based on the pervious experimental results (Fig.5-8), the generalization performance of SVM, H-ELM and LR are obviously weaker than ELM classifiers. Thus SVM, H-ELM and LR are not included in this challenge test. Different from previous experiments in which our method provides $2-3\%$ performance boost, Fig.9 and Table VI indicate that compared to the same type of features, the accuracy could be boosted to 14 percent.

F. EEG data with eye movements

Eye movement data contains heterogeneous information, such as fixation details, dispersion information, etc. The second dataset used in this study contains both EEG and eye-tracking data [2]. This dataset will be freely available to the academic community as a subset of the SEED dataset. As mentioned before, we extract DE features from EEG, and EYE movements, respectively. For DE EEG features, all the five frequency bands are used for each channel. For eye movements, we also extract DE features from five kinds of eye movement parameters: pupil diameter, dispersion, fixation duration, blink duration, and saccade. We use both linear dynamic system and moving average with the window of 20s to filter out the unrelated features for emotion recognition. Based on Table III, the total dimension of eye movement DE features for a sample is 64 ($(4+8+4+8+8)\times2$).

Then, we obtain the multiple modality DE features by combining eye movement signals and EEG data. Thus, we have two single modality DE features, and one multiple modality DE features. In other words, we have the 12 channel EEG-based DE features, the 12 channel EYE-based DE features, and the DE features combined from EYE and EEG data. Fig.10 shows the performance of single modality and multiple modality. As seen from Fig.10 (a), our learning method could provide a comparable or better performance than other classifiers. More importantly, in Fig.10 (b) the results obtained by multiple models with fusion methods outperforms the results of single modality, which shows that our method can extract more effective features from multiple modalities to enhance the emotion recognition accuracy. According to Fig.10, our learning model with 4 subnetwork nodes achieves the best performance with an average accuracy of 91.36%, which is nearly 5-10% boosted than single modality. Compared to the same type of fusion strategy-feature level fusion, our model could obtain nearly 8 percent boost.

V. Conclusion

This paper presents a hierarchical network scheme with subnetwork nodes for EEG-based emotion recognition. The problem is approached from two main directions: 1) features extracted from hundreds of network layers, rather than a single multi-layer network; 2) multiple modality features combined by early fusion. The experimental results show that our method functions as a local feature extractor and a classifier, and it performs competitively or better than other classification methods.

References

[1] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, pp. 186–197, Mar. 2010.

[2] Y. Lu, W. L. Zheng, and B. L. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pp. 1–7, 2015.

[3] G. L. Ahern and G. E. Schwartz, "Differential lateralization for positive and negative emotion in the human-brain - eeg spectral-analysis," *Neuropsychologia*, vol. 23, no. 6, pp. 745–755, 1985.

[4] W. Zheng, "Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis," *IEEE Trans. Cog. Dev. Syst.*, vol. PP, no. 99, pp. 1–1, 2016.

[5] M. Streit, J. Brinkmeyer, W. Wolwer, and W. Gaebel, "Eeg brain mapping in schizophrenic patients and healthy subjects during facial emotion recognition," *Schizophr. Res.*, vol. 61, pp. 121–122, May 2003.

[6] E. Berkman, D. K. Wong, M. P. Guimaraes, E. T. Uy, J. J. Gross, and P. Suppes, "Brain wave recognition of emotions in eeg," *Psychophysiology*, vol. 41, pp. S71–S71, 2004.

[7] P. Gifani, H. R. Rabiee, M. H. Hashemi, P. Taslimi, and M. Ghanbari, "Optimal fractal-scaling analysis of human eeg dynamic for depth of anesthesia quantification," *Journal of the Franklin Institute-engineering and Applied Mathematics*, vol. 344, pp. 212–229, May 2007.

[8] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, pp. 18–37, Jan. 2010.

[9] M. K. Kim, M. Kim, E. Oh, and S. P. Kim, "A review on the computational methods for emotional state estimation from the human eeg," *Comput. Math. Methods Med.*, p. 573734, 2013.

[10] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Trans. Affective Comput.*, vol. 5, pp. 327–339, July 2014.

[11] H. I. Suk and S. W. Lee, "A novel bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 286–299, Feb. 2013.

[12] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Eeg-based classification of music appraisal responses using time-frequency analysis and familiarity ratings," *IEEE Trans. Affective Comput.*, vol. 4, pp. 161–172, Apr. 2013.

[13] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from {EEG} data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94 – 106, 2014.

[14] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an eeg-based recognition of music liking using time-frequency analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, pp. 3498–3510, Dec. 2012.

[15] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image Vision Comput.*, vol. 31, pp. 164–174, Feb. 2013.

[16] Y. P. Lin, C. H. Wang, T. P. Jung, T. L. Wu, S. K. Jeng, J. R. Duann, and J. H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, pp. 1798–1806, July 2010.

[17] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *Int. J. Hum. Comput. Stud.*, vol. 67, pp. 607–627, Aug. 2009.

[18] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 3, pp. 162–175, 2015.

[19] R. N. Duan, J. Y. Zhu, and B. L. Lu, "Differential entropy feature for eeg-based emotion classification," *2013 6th International Ieee/embs Conference on Neural Engineering*, pp. 81–84, 2013.

[20] G. Caridakis, K. Karpouzis, and S. Kollias, "User and context adaptive neural networks for emotion recognition," *Neurocomputing*, vol. 71, pp. 2553–2562, Aug. 2008.

[21] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, "Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, pp. 589–597, May 2010.

[22] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *Int. J. Hum. Comput. Stud.*, vol. 65, pp. 724–736, Aug. 2007.

[23] S. F. Wang, Y. C. Zhu, L. H. Yue, and Q. Ji, "Emotion recognition with the help of privileged information," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, pp. 189–200, Sept. 2015.

[24] K. Takahashi and A. Tsukaguchi, "Remarks on emotion recognition from multi-modal bio-potential signals," *2003 Ieee International Conference on Systems, Man and Cybernetics, Vols 1-5, Conference Proceedings*, pp. 1654–1659, 2003.

[25] W. L. Zheng, B. N. Dong, and B. L. Lu, "Multimodal emotion recognition using eeg and eye tracking data," in *Proc. IEEE Int. Conf. Engin. Medicine & Biology Society*, pp. 5040–5043, 2014.

[26] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Net.*, vol. 61, pp. 85–117, Jan. 2015.

[27] J. Weng, N. Ahuja, and T. S. Huang, "Cresceptron: a self-organizing neural network which grows adaptively," in *Proc. Int. Jt. Conf. Neural. Netw.*, vol. 1, (Baltimore, US.), pp. 576–581, Jun. 1992.

[28] K. Fukushima, "Neocognitron: A self-organizing neural network for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.

[29] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, 1994.

[30] N. Schraudolph and T. J. Sejnowski, "Unsupervised discrimination of clustered data via optimization of binary information gain," in *Proc. Adv. Neural Inf. Process. Syst.*, (San Mateo, US.), pp. 499–506, 1993.

[31] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Back-propagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[32] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, July 2006.

[33] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, (Vancouver, BC, Canada), 2007.

[34] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec 2010.

[35] L. L. C. Kasun, H. Zhou, G.-B. Huang, and C.-M. Vong, "Representational learning with extreme learning machine for big data," *IEEE Intell. Syst.*, vol. 28, pp. 31–34, Nov 2013.

[36] M. Chen, K. Weinberger, Z. Xu, and F. Sha, "Marginalizing stacked autoencoders," *J. Mach. Learn. Res.*, vol. 22, no. 2, pp. 191–194, 2015.

[37] J. Cao, Y. Zhao, X. Lai, M. E. H. Ong, C. Yin, Z. X. Koh, and N. Liu, "Landmark recognition with sparse representation classification and extreme learning machine," *J. Franklin Inst.*, vol. 352, pp. 4528–4545, July 2015.

[38] Y. Yang and Q. M. J. Wu, "Multilayer extreme learning machine with subnetwork nodes for representation learning," *IEEE Trans. Cybern.*, vol. 46, pp. 2570–2583, Nov 2016.

[39] J. Cao, K. Zhang, M. Luo, C. Yin, and X. Lai, "Extreme learning machine and adaptive sparse representation for image classification," *Neural Networks*, vol. 81, pp. 91 – 102, 2016.

[40] W. L. Zheng, J. Y. Zhu, Y. Peng, and B. L. Lu, "Eeg-based emotion classification using deep belief networks," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2014.

[41] R. Khosrowabadi, C. Quek, K. K. Ang, and A. Wahab, "Ernn: A biologically inspired feedforward neural network to discriminate emotion from eeg signal," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 25, pp. 609–620, Mar. 2014.

[42] S. Fusi, E. K. Miller, and M. Rigotti, "Why neurons mix: high dimensionality for higher coginition," *Curr. Opin. Neurobiol.*, vol. 37, pp. 66–74, April 2016.

[43] M. Rigotti, O. Barak, M. R. Warden, X. J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," *Nature*, vol. 497, pp. 585–590, May 2013.

[44] M. Lin, Q. Chen, and S. Yan, "Network in network," *Arvix: 1312.4400*, 2013.

[45] Y. Yang and Q. M. J. Wu, "Extreme learning machine with subnetwork hidden nodes for regression and classification," *IEEE Trans. Cybern.*, vol. 46, pp. 2885–2898, Dec 2016.

[46] G.-B. Huang, Z. Bai, L. Lekamalage, C. Kasun, and C. M. Vong, "Local receptive fields based extreme learning machine," *IEEE Comput. Intell. Mag.*, vol. 10, pp. 18–29, May 2015.

[47] J.-X. Tang, C.-W. Deng, and G.-B. Huang, "Extreme Learning Machine for Multilayer Perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, pp. 809–821, April 2015.

[48] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space," *IEEE Trans. Affective Comput.*, vol. 2, pp. 192–205, Oct. 2011.

[49] Y. Dong, S. Gao, K. Tao, J. Q. Liu, and H. L. Wang, "Performance evaluation of early and late fusion methods for generic semantics indexing," *Pattern Analysis and Applications*, vol. 17, pp. 37–50, Feb. 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2685338, IEEE Transactions on Cognitive and Developmental Systems

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, June 2015.
[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.
[52] G.-B. Huang, H.-M. Zhou, X.-J. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man. Cy. B.*, vol. 42, pp. 513–529, April 2012.
[53] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[54] T. Murofushi and M. Sugeno, "Affectively intelligent and adaptive car interfaces," *Inofmration Science*, vol. 29, no. 2, pp. 201–227, 1989.

Yimin Yang (S'10-M'13) received his Ph.D. degrees in electrical engineering from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2013.

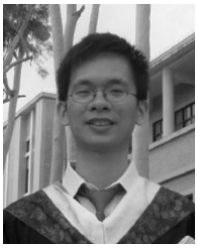
He is currently a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering at the University of Windsor, ON, Canada. He has authored or coauthored more than 30 refereed papers. His research interests are artificial neural networks, feature extraction, and sensor analysis and fusion.

Dr. Yang was the recipient of the Outstanding Ph.D. Thesis Award of Hunan Province, and the Outstanding Ph.D. Thesis Award Nominations of Chinese Association of Automation, China, in 2014 and 2015, respectively. He has been serving as a reviewer for international journals of his research field, such as the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Cybernetics, etc.

Q. M. Jonathan Wu (M'92-SM'09) received his Ph.D. in electrical engineering from the University of Wales, Swansea, U.K., in 1990.

He was affiliated with the National Research Council of Canada for ten years beginning in 1995, where he became a Senior Research Officer and a Group Leader. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He is a Visiting Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He has published more than 300 peer-reviewed papers in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include 3-D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks.

Dr. Wu holds the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems. He is an Associate Editor of the IEEE Transactions on Neural Networks and Learning Systems, and the Cognitive Computation. He has served on technical program committees and international advisory committees for many prestigious conferences.



Wei-Long Zheng (S'14) received the bachelor's degree in information engineering from the Department of Electronic and Information Engineering, South China University of Technology, Guangzhou, in 2012

He is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His research focuses on affective computing, brain-computer interface, machine learning, and pattern recognition

Bao-Liang Lu (M'94-SM'01) received the B.S. degree in instrument and control engineering from Qingdao University of Science and Technology, Qingdao, China, in 1982, the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xian, China, in 1989, and the Dr. Eng. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1994.

He was with Qingdao University of Science and Technology from 1982 to 1986. From 1994 to 1999, he was a Frontier Researcher with the Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan, from 1999 to 2002. Since 2002, he has been a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He has also been an Adjunct Professor with the Laboratory for Computational Biology, Shanghai Center for Systems Biomedicine, since 2005. His current research interests include brain-like computing, neural network, machine learning, computer vision, bioinformatics, brain-computer interface, and affective computing.

Dr. Lu was the President of the Asia Pacific Neural Network Assembly (APNNA) and the General Chair of the 18th International Conference on Neural Information Processing in 2011. He is currently an Associate Editor of the IEEE Transactions on Cognitive and Developmental Systems, and the Neural Networks.