Spatial–Temporal Recurrent Neural Network for Emotion Recognition

Tong Zhang, Wenming Zheng, *Member, IEEE*, Zhen Cui, Yuan Zong, and Yang Li

***Abstract*—In this paper, we propose a novel deep learning framework, called spatial–temporal recurrent neural network (STRNN), to integrate the feature learning from both spatial and temporal information of signal sources into a unified spatial–temporal dependency model. In STRNN, to capture those spatially co-occurrent variations of human emotions, a multidirectional recurrent neural network (RNN) layer is employed to capture long-range contextual cues by traversing the spatial regions of each temporal slice along different directions. Then a bi-directional temporal RNN layer is further used to learn the discriminative features characterizing the temporal dependencies of the sequences, where sequences are produced from the spatial RNN layer. To further select those salient regions with more discriminative ability for emotion recognition, we impose sparse projection onto those hidden states of spatial and temporal domains to improve the model discriminant ability. Consequently, the proposed two-layer RNN model provides an effective way to make use of both spatial and temporal dependencies of the input signals for emotion recognition. Experimental results on the public emotion datasets of electroencephalogram and facial expression demonstrate the proposed STRNN method is more competitive over those state-of-the-art methods.**

***Index Terms*—Electroencephalogram (EEG) emotion recognition, emotion recognition, facial expression recognition, spatial–temporal recurrent neural network (STRNN).**

I. Introduction

HUMAN emotion analysis plays a crucial role in endowing artifact machines with humanized characteristics, and is arousing more and more attentions due to its potential applications to human–machine interaction. For example, humanoid robots have been employed in families or shops, which are capable of recognizing the principal human emotions and adapting their behavior to the mood of their interlocutors. Despite the successfully practical application, the study of emotion analysis is still facing significant difficulties due to its intrinsic property of less tangibility. To solve this problem, researchers utilize different electric devices to obtain emotions conveyed in external signal forms. With the development of hardware techniques, it is becoming easy to collect signals reflecting human emotions, such as acoustic waves, facial video sequences, electroencephalogram (EEG) signals, and so on.

Among various emotion signals, EEG and facial expression sequences are widely employed for emotion analysis. For them, time-varying signals are either produced from multiple active electrodes attached on cerebral cortex by arranging a certain spatial layout, or collected by using general cameras. These emotion related signals contain not only spatial components at a single moment but also contextual dependencies among temporal slices. In order to better recognize human emotion, the crucial spatial, and temporal dependencies should be well modeled. Existing methods, such as deep belief networks (DBNs) [2], canonical correlation analysis (CCA) [33], and so on, have been proposed to deal with EEG signals which mainly focus on capturing spatial correlations among electrodes. For facial expression recognition, the algorithm proposed in [12] aims to model spatial and temporal relationship by combining convolutional neural network (CNN) [3], [4] (capturing only spatial information) and recurrent neural network (RNN) [5] (capturing only temporal information). Although these algorithms have achieved notable performance, they are still necessary to be improved.

1) The complex spatial dependencies between multiple electrodes or facial local areas should be well modeled. The existing frameworks, e.g., CNN in [12], may fail to capture long-distance spatial dependencies due to the locality of convolution and pooling layers.
2) Temporal variation in EEG signals is also important for emotion recognition. However, existing algorithms [2], [9], [11] just focus on dealing with EEG signals with a temporal slice ignoring the temporal variation.

In this paper, we propose a unified deep network framework called spatial–temporal RNN (STRNN) to deal with both EEG-based emotion recognition and facial emotion recognition. STRNN can not only learn spatial dependencies of multielectrode or image context itself, but also learn a

Manuscript received June 26, 2017; revised September 30, 2017; accepted December 18, 2017. Date of publication January 30, 2018; date of current version February 14, 2019. This work was supported in part by the National Basic Research Program of China under Grant 2015CB351704, in part by the National Natural Science Foundation of China (NSFC) under Grant 61231002, Grant 61602244, and Grant 61772276, and in part by the Key Research and Development Program of Jiangsu Province under Grant BE2016616. This paper was recommended by Associate Editor B. W. Schuller. *(Corresponding author: Wenming Zheng.)*

T. Zhang and Y. Li are with the Key Laboratory of Child Development and Learning Science, Ministry of Education, and the Department of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: tongzhang@seu.edu.cn; yang_li@seu.edu.cn).

W. Zheng and Y. Zong are with the Key Laboratory of Child Development and Learning Science, Ministry of Education, Research Center for Learning Science, Southeast University, Nanjing 210096, China (e-mail: wenming_zheng@seu.edu.cn; xhzongyuan@seu.edu.cn).

Z. Cui is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zhen.cui@njust.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2788081

long-term memory information in temporal sequences. To learn spatial dependencies, a quad-directional spatial RNN (SRNN) layer is first employed to scan each temporal slice by adopting certain spatial orders, respectively, from different angles, and finally produce a discriminative dependency sequence in the slice. Comparing to CNN for modeling spatial information, RNN is advantageous in capturing long-distance spatial dependencies and modeling the relationship between two areas which are far away from each other. Then, a bi-directional temporal RNN (TRNN) layer is further stacked on SRNN to capture long-term temporal dependencies by scanning the temporal sequences forward and backward. In each RNN layer, the previous states are connected to the current one so that the sub-network is inherently deep and able to retain all the past inputs. The benefit of the hierarchical RNN is that the two layers may act as two memory units to remember and encode all the scanned spatial and temporal area so that the proposed STRNN is able to globally model the spatial and temporal dependencies.

As emotion stimuli are usually activated in some local regions, we expect to discover those salient stimulus regions expressing human emotions. To this end, we introduce the sparse projection transformation onto those spatially encoding states to detect those salient activation points. Besides, as the global projection is operated on the entire spatial domain, the learned STRNN can automatically bundle those co-occurrent emotion regions, which would bring some gains for the final emotion recognition. Similarly, the sparse projection is also used for those temporally encoded states in order to adaptively choose and combine those time slices with more discriminability.

The main contributions of this paper can be summarized as follows.

1) We propose an end-to-end STRNN to jointly integrate spatial and temporal dependencies, as well as learn discriminative features. To the best of our knowledge, few models in the previous work utilize RNNs to capture both spatial and temporal information.
2) Two emotion recognition tasks, i.e., EEG-based emotion recognition and facial emotion recognition, are investigated and unified under a deep network framework by constructing spatial–temporal volumes, where EEG signals are spatially organized in electrode coordination.
3) Salient emotion activation regions can be effectively captured by introducing sparse projection on those encoding hidden states, which can naturally bundle those co-occurred emotion activation regions by adaptively weighting them.

This paper is organized as follows. Section II overviews some approaches related to RNN and human emotion recognition in EEG signals and videos, in Section III we proposed the STRNN method for emotion recognition in details, we present experimental results to evaluate the proposed method in Section IV and finally in Section V we conclude this paper.

II. Related Work

Human emotion recognition based on EEG signals or facial expression sequences had been extensively investigated during the past decades, and a lot of algorithms have been proposed in the literatures to this end. For instances, for EEG-based emotion recognition, descriptors, such as high order crossings [6] and differential entropy (DE) [7] are employed, and popular classifiers, such as support vector machine (SVM) [8] and group sparse CCA (GSCCA) [11] are used to achieve classification. Recent years, deep neural networks including DBNs [2] and bimodal deep autoencoder (BDAE) [10] are also used to learn high-level features from the extracted descriptors, which achieve competitive performance.

On the other hand, for facial expression sequences-based emotion recognition, various hand-crafted facial features has been applied, such as 3-D HOG [13], 3-D SIFT [17], expressionlet-based spatio-temporal manifold (STM-ExpLet) [18], and so on. In addition, deep learning methods are also employed to deal with the expression recognition problem from facial image sequences in recent years, e.g., the 3DCNN method proposed in [19]. The method named 3DCNN-DAP proposed in [19] employs 3DCNN while using the strong spatial structural constraints on the dynamic action parts to extract robust representation from videos.

Recently, RNNs have achieved great success in processing sequential data, such as natural language processing [20], [21], action recognition [25], speech recognition [26], and so on. Then RNN is further improved to deal with images [27] by scanning the regions of images into sequences in certain directions. Due to the ability of retaining information about the past inputs, RNN is able to learn contextual dependencies with images, which is advantageous compared to CNN. This is because CNN may fail to learn the global dependencies due to the locality of convolution and pooling layers. Moreover, RNN is able to model temporal dependencies and can be used as a complement to CNN which captures only spatial dependencies. For this reason, RNN is usually combined with CNN to model spatio-temporal information in some recognition tasks [22]–[24], [28], [29].

In what follows, we will propose our STRNN method, which uses an end-to-end spatial–temporal learning network to simultaneously integrate spatial and temporal dependencies for co-adaptively dealing with EEG-based emotion recognition and video-based facial expression recognition.

III. STRNN Model for Emotion Recognition

To specify the proposed STRNN method, we illustrate the framework of STRNN in Fig. 1, in which the inputs of networks would be any spatial–temporal style data, such as multichannel sequence signals (e.g., EEG) or spatial–temporal cubic volumes (e.g., videos) as long as they could be traversed in certain orders in space or time domain. In Fig. 1, we take cubic videos as an example for the simplicity of description. The goal of the proposed framework is to capture spatial–temporal information within sequence signals. To realize this point, we attempt to stack two-layer RNNs, i.e., an SRNN layer and a TRNN layer, so as to concatenating other layers for an end-to-end deep neural network. Consequently, STRNN combines spatial and temporal information simultaneously by building the dependencies of adjacent and even

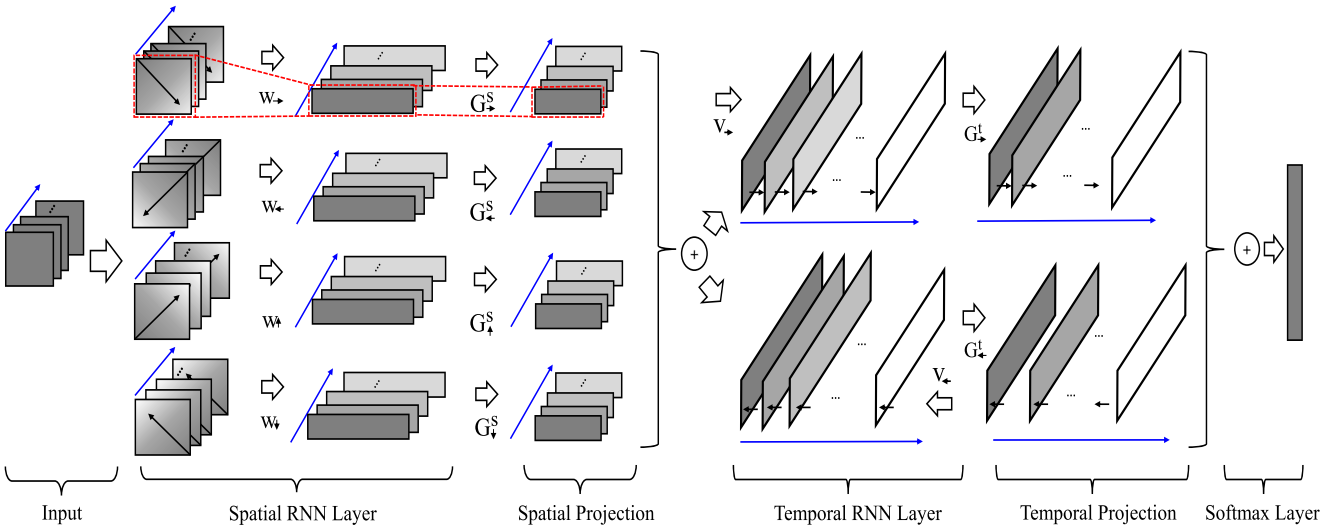

Fig. 1. Overview on the proposed STRNN framework. The SRNN and TRNN are elaborately integrated and jointly learned to capture spatial and temporal dependencies. The blue arrow indicates the temporal axis. More details can be found in Section III.

long-term elements. Moreover, to detect those salient emotion regions, sparse projections are further applied to those encoding hidden states of SRNN and TRNN layers.

To model spatial dependencies on each time slice, i.e., the relationships between responses of multiple electrodes at a certain moment for EEG signals, we use RNN to scan all spatial elements under a predefined order strategy. Unlike those Markov chain structures frequently used for such a graph model, RNN simplifies its process by unfolding graph structures into an order structure, which makes the learning more controllable. After scanning the time slice element by element sequentially, RNN can characterize those low-level elements and their complex dependencies if long-term recurrent units (e.g., long short-term memory, LSTM [30]) are adopted. However, it is notable that the data could be contaminated by signal noises in EEG or partial occlusion in videos, a single RNN used in a 2-D space may not be enough to resist these variations. For this problem, we use four directional RNNs to traverse the spatial region at a time slice from four specific angles. The four RNNs are actually complementary for constructing a complete relationship graph, and thus alleviate the effect of noises while simplifying those techniques used for modeling graph structures. Concretely, when modeling spatial dependencies, we use a graph $\mathcal{G}_t = \{\mathcal{R}_t, \mathcal{C}_t\}$ to represent the spatial elements in the tth slice denoted by $\mathbf{X}_t$, where $\mathcal{R}_t = \{\mathbf{x}_{tij}\}(i = 1, \ldots, h, j = 1, \ldots, w)$ represents the vertex set of spatial elements indexed by their spatial coordinates, and $\mathcal{C}_t = \{e_{tij,tkl}\}$ denotes the edges of spatial neighboring elements in the tth slice. Then we traverse through $\mathcal{G}_t$ with a predefined forward evolution sequence so that the input state and the previous states can be defined for a RNN unit. Formally, the adopted multidirectional SRNNs in STRNN can be written as

$$\mathbf{h}^r_{tij} = \sigma_1\left(\mathbf{U}^r\mathbf{x}_{tij} + \sum_{k=1}^{h}\sum_{l=1}^{w}\mathbf{W}^r\mathbf{h}^r_{tkl} \times e_{tij,tkl} + \mathbf{b}^r\right) \quad (1)$$

$$e_{tij,tkl} = \begin{cases} 1, & \text{if } (k,l) \in \mathcal{N}^r_{ij} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{x}_{tij}$ and $\mathbf{h}^r_{tij}$, respectively, denote the representation of input and hidden node at the location of (i, j) in the tth slice. $\mathcal{N}^r_{ij}$ is the set of predecessors of the vertex (i, j) and r represents a certain traversing direction. Concretely, for example, for the directional traversing from the top-left corner, $\mathcal{N}^r_{ij}$ is defined as $\{(i, i-1), (i-1, j-1), (i-1, j)\}$. $\mathbf{U}^r, \mathbf{W}^r, \mathbf{b}^r$ are learnable parameters in SRNNs and the nonlinear function denoted by $\sigma_1(\cdot)$ for hidden layers is ReLU or Sigmoid function. Hence, $\mathbf{h}^r_{tij}$ collects information of all the previous scanned elements of the current state (i, j).

As SRNN layer traverses all the vertexes in $\mathcal{R}_t$, the number of the hidden states equals $h \times w$ for a given traversing direction. For simplification, the output hidden states denoted as $\mathbf{h}^r_{tij}(i = 1, \ldots, h, j = 1, \ldots, w)$ are rewritten as $\mathbf{h}^r_{tk}(k = 1, \ldots, K)$, where K equals $h \times w$. To further detect those salient regions of emotion representation, projection matrices are applied to the spatial hidden states corresponding to different traversing directions. Assume that a projection matrix for a certain traversing direction is denoted as $\mathbf{G}^r = [G^r_{ij}]_{K \times K_d}$ where K_d denotes the number of the hidden states after projection, then the projection can be written as

$$\mathbf{s}^r_{tl} = \sum_{i=1}^{K} G^r_{il}\mathbf{h}^r_{ti}, l = 1, \ldots, K_d \quad (3)$$

where $\mathbf{s}^r_{tl}$ denotes the lth output feature vector after projection. Let $\mathbf{s}^r_t = [(\mathbf{s}^r_{t1})^T, \ldots, (\mathbf{s}^r_{tK_d})^T]^T$ denotes the concatenated feature vector of a certain traversing direction, then the output of SRNN layer summarizes the stimulus from all directions $\mathcal{D}$

$$\mathbf{m}_t = \sum_{r \in \mathcal{D}} \mathbf{P}^r\mathbf{s}^r_t \quad (4)$$

where $\mathbf{P}^r$ is the learnable matrix corresponding to each direction. Until now, for an input slice $\mathbf{X}_t$, the output feature has been generated which is denoted as $\mathbf{m}_t$. Such a process is designed as a network layer called SRNN layer to seamlessly connect other layers.

In SRNN layer, we consider four traversing directions starting from four angular points to make the traversing information from traversing processes mutually complementary. For example, the directional traversing from the top-left corner aims to capture contextual cues about the top-left areas with the adjacent predecessor set. Thus, four directed acyclic chains can be generated to represent the 2-D neighborhood system by connecting contiguous elements and traversing these elements, respectively, from four directions. By doing this, discriminative spatial dependencies for emotion recognition can be modeled.

The representations learned from the SRNN layer are sequentially concatenated at each time slice and thus form a temporal sequence. For an entire emotion process, a single slice cannot reflect the characteristic of emotion due to its small granularity. The better strategy is to build the entire dynamic process rather than isolating considering each slice. RNNs can adaptively model such a temporal dynamic process. Here, we employ a bi-directional RNN to simultaneously capture forward and backward dynamic transforms of sequence, i.e., two RNNs are, respectively, used to traverse the temporal sequence in a forward or backward behavior. Formally, suppose that sequential representations are denoted as $\mathbf{m}_t$ and the temporal length is L, then the TRNN layer can be written as

$$\mathbf{h}_t^f = \sigma_1\left(\mathbf{R}^f\mathbf{m}_t + \mathbf{V}^f\mathbf{h}_{t-1}^f + \mathbf{b}^f\right) \tag{5}$$

$$\mathbf{h}_t^b = \sigma_1\left(\mathbf{R}^b\mathbf{m}_t + \mathbf{V}^b\mathbf{h}_{t-1}^b + \mathbf{b}^b\right) \tag{6}$$

where $\{\mathbf{R}^f, \mathbf{V}^f, \mathbf{b}^f\}$ and $\{\mathbf{R}^b, \mathbf{V}^b, \mathbf{b}^b\}$ are the learnable parameters for recurrently traversing the sequences scanned forward and backward, respectively, $\mathbf{m}_t$, $\mathbf{h}_t^f$, and $\mathbf{h}_t^b$ are the input nodes, hidden nodes for the forward scanned network and hidden nodes the backward scanned network, respectively. Similar to spatial projection in SRNN layer, projection matrices denoted as $\mathbf{G}^f = [G_{ij}^f]_{L\times L_d}$ and $\mathbf{G}^b = [G_{ij}^b]_{L\times L_d}$ are also applied to detecting the salient temporal hidden states, resulting in the following expressions:

$$\mathbf{q}_t^f = \sum_{i=1}^{L} G_{it}^f\mathbf{h}_i^f, \quad \mathbf{q}_t^b = \sum_{i=1}^{L} G_{it}^b\mathbf{h}_i^b, \quad t = 1, \ldots, L_d \tag{7}$$

where L_d denotes the sequence length after temporal projection and $\mathbf{q}_t^f$, $\mathbf{q}_t^b$, respectively, denote the tth output feature vectors of the forward and backward scanned networks.

Let

$$\mathbf{q}^f = \left[\left(\mathbf{q}_1^f\right)^T, \ldots, \left(\mathbf{q}_{L_d}^f\right)^T\right]^T$$

and

$$\mathbf{q}^b = \left[\left(\mathbf{q}_1^b\right)^T, \ldots, \left(\mathbf{q}_{L_d}^b\right)^T\right]^T$$

denote the concatenated vectors for the forward and backward scanned networks, respectively, then, the output of TRNN layer denoted as $\mathbf{o}$ can be calculated by the following equation:

$$\mathbf{o} = \mathbf{P}^f\mathbf{q}^f + \mathbf{P}^b\mathbf{q}^b \tag{8}$$

where $\mathbf{o} = [o_1, o_2, \ldots, o_C]^T$ and C equals the number of emotion types.

Finally, the output nodes of TRNN layer are fed into the softmax layer for emotion classification

$$P(i|\mathbf{X}) = \exp(o_i)\Big/\sum_{k=1}^{C}\exp(o_k) \tag{9}$$

where $P(i|\mathbf{X})$ denotes the probability for the input $\mathbf{X}$ being predicted as the ith class.

In addition, we use cross entropy loss defined as follows to represent the objective loss function, which can be written as:

$$E = -\sum_{i=1}^{N}\sum_{c=1}^{C}\tau(y_i, c)\times\log P\left(c|\mathbf{X}^i\right) + \lambda_1\sum_{r\in\mathcal{D}}\sum_{i=1}^{K_p}\left\|\mathbf{g}_i^r\right\|_1 + \lambda_2\left(\sum_{i=1}^{L_p}\left\|\mathbf{g}_i^f\right\|_1 + \sum_{i=1}^{L_p}\left\|\mathbf{g}_i^b\right\|_1\right) \tag{10}$$

in which

$$\tau(y_i, c) = \begin{cases} 1, & \text{if } y_i = c \\ 0, & \text{otherwise} \end{cases}$$

where E denotes the cross entropy loss, N denotes the number of the training samples, $\mathbf{X}^i$ represents the ith training sample of the training set, y_i is the label of the ith training sample, $\mathbf{g}_i^r$, $\mathbf{g}_i^f$, and $\mathbf{g}_i^b$ denote the ith column vectors of $\mathbf{G}^r$, $\mathbf{G}^f$, and $\mathbf{G}^b$, respectively.

In the loss function, the first term calculates the mean negative logarithm value of the prediction probability of the training samples. The second and third terms ensure the sparse structure of the matrices in spatial and temporal projection. As elements of projection matrices indicate the importance of the corresponding spatial or temporal hidden states, the sparse structure is able to endow high weights to the discriminative hidden states while low weights to others, which achieves the purpose of selecting salient hidden states.

The proposed STRNN can be effectively optimized by the classic back propagation through time (BPTT) algorithm. In BPTT, the recurrent nets can be converted into common feedforward networks after they are unfolded to a sequence with a limited size. Thus, traditional gradient back-propagation used in common deep networks can be directly applied.

IV. Experiments

In this section, we first introduce the datasets we use for testing the performance of our proposed STRNN, then report and analyze the results of our method on these datasets by comparing with other state-of-the-art methods.

A. Datasets and Feature Extraction

The proposed STRNN method is tested on both SJTU emotion EEG dataset (SEED) [2] and the dataset of CK+ facial expression image sequences [31]. SEED contains three categories of emotions (positive, neutral, and negative) of fifteen subjects (seven males and eight females), which are elicited by showing emotional film clips to the participants. The EEG signals of these subjects were recorded using an ESI neuroscan system at a sampling rate of 1000 Hz from 62-channel

Fig. 2. Samples of data augmentation of CK+. The first row contains four original frames sampled from a sequence, and the second and third rows contain the images which are rotated 7° clockwise and 12° counterclockwise corresponding to the images in the first row.

electrode cap according to the international 10–20 system. The CK+ dataset consists of 327 image sequences with seven emotion labels: 1) anger (An); 2) contempt (Co); 3) disgust (Di); 4) fear (Fe); 5) happiness (Ha); 6) sadness (Sa); and 7) surprise (Su) of 118 subjects. In this database, each sequence starts with a neutral emotion and ends with a peak of the emotion.

To recognize emotion from EEG signals, DE descriptors [2] are extracted, which are calculated in five frequency bands (delta: 1–3 Hz, theta: 4–7 Hz, alpha: 8–13 Hz, beta: 14–30 Hz, and gamma: 31–50 Hz) of 62 channels. For a specified continuous EEG sequence, a 256-point short-time Fourier transform with a nonoverlapped Hanning window of 1s is used to extract five frequency bands of EEG signals and DE is calculated for each frequency band. After this process, discrete sequences in five bands of 62 channels are generated. Then we use a slicing window of 9s to temporally scan the sequences by one step. For each step, the sequences in the slicing window are used as the representation of the point which is in the center of the slicing window. By doing this, the temporal dependencies can be involved while recognizing the human emotion at a specific moment. This is quite different from [2] which just focuses on recognizing the average energy within a short time ignoring the temporal variation information. For CK+, we use the pretrained model proposed in [4] to extract features in each image to reduce the effects of noises or variant face poses so as to improve the representation ability. As the number of the training samples is limited, we perform rotation transformation to the sequences to achieve data augmentation. Each image is rotated with angles including 7°, −7°, 12°, and −12° so that there are totally 1635 samples. Some examples of this rotation process are shown in Fig. 2. For each image, the feature maps of the pooling layer, which is located before the first fully connected layer of the pretrained model, are used as the representation. For each input image sequence, the extracted feature maps are concatenated temporally as the representation of this sequence.

B. EEG Signals-Based Emotion Recognition

The basic experiment configuration is the same to the one in [2]. In this dataset, there are totally fifteen subjects and each subject is conducted with the experiments across two time sessions. Thus, there are totally 30 experiments evaluated here. Following the same protocol in [2], the training data and the testing data are, respectively, taken from different sessions of the same experiment. There are nine sessions for training and the remaining six sessions for testing.

In SRNN layer, the numbers of the input, hidden, and output nodes are set to be 5, 30, and 30, respectively, and the number of hidden states [K_p in (3)] is reduced from 62 to 10 after spatial projection. In TRNN layer, the numbers of hidden and output nodes are set to be 30 and 3, where the number of the output nodes is set according to the number of emotion types. The number of hidden states [L_p in (7)] is reduced from 9 to 5 after temporal projection. These parameters of our STRNN are roughly set without elaborate traversal. In SRNN layer, the RNNs scan the electrodes from four angels. As the distribution of locations of the electrodes is not exactly a rectangle, we define the scanning order as shown in Fig. 3 to model intimate interactions existing among those spatially adjacent electrodes.

The average accuracy of STRNN with the DE features of all frequency bands in 30 experiments of fifteen subjects is shown in Table I. This result is also compared with various existed algorithms including linear discriminant analysis (LDA) [32], CCA, graph regularized sparse LDA (GraphSLDA) [9], BDAE [10], DBN, and so on. Most of these methods employ DE features of all 62 channels except SVM [2], which uses both 62 and 12 channels. SVM of 62 channels achieves the accuracy of 83.99% and its performance is further improved to be 86.65% by selecting certain 12 channels out of full 62 channels. CCA gets the performance of 76.16% while GSCCA [11] achieves much higher accuracy of 82.35% by endowing conventional CCA with the ability of handling the group feature selection problem from raw EEG features. BDAE seems to achieve higher performance comparing to our STRNN. However, it follows a different experimental protocol which uses 27 data files from only 9 subjects instead of full 15 subjects. Moreover, it also employs additional information of eye movement together with DE features. GraphSLDA achieves the accuracy of 88.41% with a leave-one-session-out protocol which uses 14 sessions for training and the left one for testing. Our STRNN achieves the accuracy of 89.50% using data files of 15 subjects, which is highest among those algorithms following the same protocol. This performance gain indicates that our STRNN benefits from modeling the spatial and temporal dependencies layer by layer while the comparison algorithms do not consider the spatio-temporal structure of the EEG signals.

To reveal which frequency oscillation of brain activity is more related to the emotion processing, the performance of the DE feature on different frequency bands (Delta, Theta, Alpha, Beta, and Gamma) are compared between DBN and STRNN, which is shown in Table II. As we can see, the distribution of the accuracies of STRNN on different frequency bands is quite different from the result of DBN: the accuracies achieved on four frequency bands (Delta, Theta, Alpha, and Beta) are all more than 80% while the accuracy of Gamma is lower. The highest accuracy is achieved on Beta band. However, for the results of DBN in [2], only beta and gamma bands of

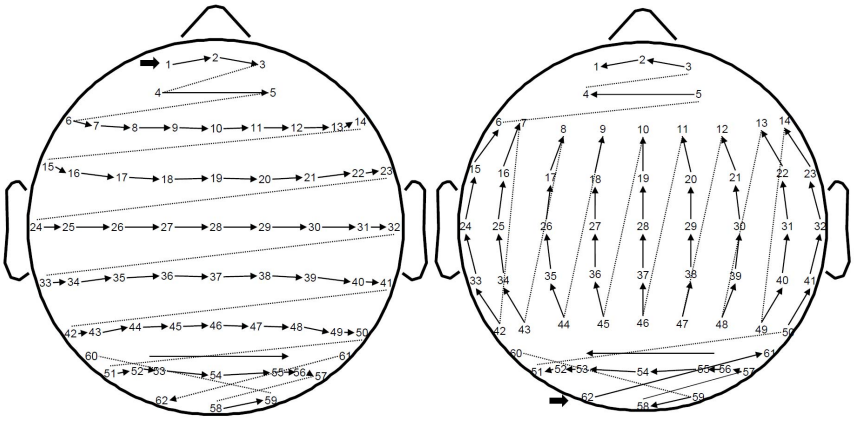

Fig. 3. Scanning order of electrodes in two directions. The other two scanning directions inverse the current scanning orders.

TABLE I
COMPARISONS ON EEG SIGNAL-BASED EMOTION DATASET SEED

Feature	classifier	channels number	frequency bands number	number of subjects	training session / testing session	accuracy (%)
DE	SVM [2]	62	5	15	9/6	83.99
DE	SVM [2]	12	5	15	9/6	86.65
DE	LDA [32]	62	5	15	9/6	80.32
DE	CCA [33]	62	5	15	9/6	76.16
DE and eye movement	BDAE [10]	62	5	9	9/6	91.01
DE	GraphSLDA [9]	62	5	15	14/1	88.41
DE	GSCCA [11]	62	5	15	9/6	82.35
DE	DBN [2]	62	5	15	9/6	86.08
DE	STRNN	62	5	15	9/6	**89.50**

TABLE II
PERFORMANCE OF DIFFERENT FREQUENCY BANDS ON SEED

frequency band	Delta	Theta	Alpha	Beta	Gamma	all
DBN [2]	64.32/12.45	60.77/10.42	64.01/15.97	78.92/12.48	**79.19/14.58**	86.08/8.34
STRNN	**80.9/12.27**	**83.35/9.15**	**82.69/12.99**	**83.41/10.16**	69.61/15.65	**89.50/7.63**

EEG signals are more related with emotion processing than other frequency bands. This difference may be caused by the fact that the temporal slicing window we use during the feature extraction process involves temporal dependencies. And according to our results, the spatial–temporal dependencies of the four frequency bands, i.e., Delta, Theta, Alpha, and Beta, contribute more to the recognition of emotion. Moreover, the deviations of recognition results of all five bands are calculated as well as those of each specific frequency band. The values of deviations of STRNN are lower than DBN except Gamma band, which indicates the performance of our STRNN is more stable across different experiments of different subjects.

Fig. 4 shows the confusion matrix of all evaluated experiments for SEED, where the element located in the *i*th row and *j*th column means the percentage of those samples which belong class *i* and are predicted as class *j*. As it is shown, our algorithm performs well in recognizing all three types of emotions as the accuracies of them are more than 85.0%. Positive and negative emotions are easier to be recognized whereas neutral, by contrast, is relatively difficult to be correctly classified as it is easily confused with negative.

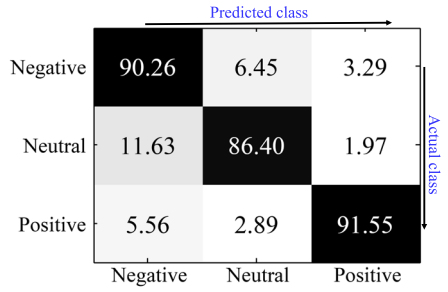

Fig. 4. Experimental results of confusion matrix on SEED.

C. Video Emotion Recognition

Following the previous protocol [18], [37], we train and test the CK+ database with the subject independent cross

TABLE III
COMPARISONS ON VIDEO FACE-BASED EMOTION DATASET CK+

method	cross validation protocol	data type	accuracy(%)
3D SIFT [17]	10-fold	video	81.4
3D HOG [13]	10-fold	video	91.4
MSR [34]	10-fold	video	91.4
Cov3D [35]	5-fold	video	92.3
TMS [36]	4-fold	video	91.9
C_{nclass} [14]	10-fold	static image	92.1
C_{bin} [14]	10-fold	static image	96.2
F-Bases [15]	10-fold	video	92.6
DNN [16]	10-fold	static image	90.91
3DCNN [19]	10-fold	video	85.9
3DCNN-DAP [19]	10-fold	video	92.4
STM-ExpLet [18]	10-fold	video	94.2
DTAN [37]	10-fold	video	91.4
STRNN	10-fold	video	**95.4**

validation in this experiment to achieve a fair evaluation. In this process, the database is divided into ten groups according to the subject ID and there are no overlapping samples among these ten groups. Then experiments are conducted on these ten divided groups with ten runs in total. For each run, nine groups were employed for training and the remaining one group for testing. Such ten runs are performed by enumerating the group used for testing and the average recognition performance is computed as the final result of the ten runs.

The parameters of our STRNN are set as follows. The numbers of the input, hidden, and output nodes in SRNN layer are set to be 512, 50, and 50, respectively, the number of hidden states [K_p in (3)] is reduced from 49 to 10 after spatial projection. In TRNN layer, the numbers of hidden and output nodes are set to be 150 and 7. The number of hidden states [L_p in (7)] is reduced from 44 to 5 after temporal projection.

As it is shown in Table III, various state-of-the-art methods for CK+ are compared with our algorithm, including Cov3D [35], TMS [36], STM-ExpLet [18], and so on. Most of these methods take videos as input and adopt tenfold cross validation as same as our STRNN, while the others, e.g., TMS [36], C_{bin} [14], and DNN [16], adopt different experimental settings by either taking static images as input or using different cross validation protocols. For a fair comparison, we mainly focus on those algorithms with the same experimental settings. The hand-drafted feature-based algorithms, including 3-D SIFT [17], 3-D HOG [13], MSR [34], and F-Bases [15], achieve the accuracies of 81.4%, 91.4%, 91.4%, and 92.6%, respectively. Higher accuracy of 94.2% is achieved by STM-ExpLet which introduces complex manifold structures. Besides, various deep learning methods are also applied to recognize facial expressions including 3-D CNN, 3DCNN-DAP, deep temporal appearance network (DTAN) [37], and so on. 3-D CNN only gets the accuracy of 85.9% while

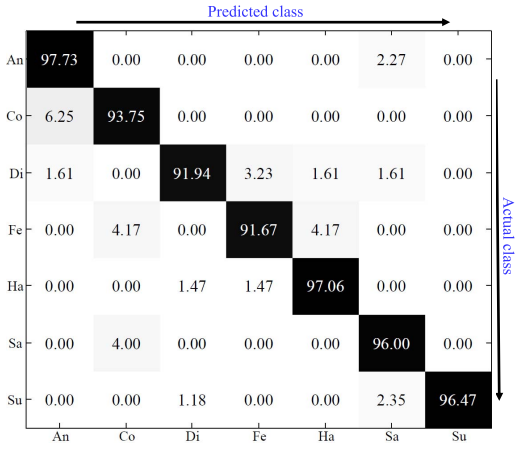

Fig. 5. Experimental results of confusion matrix on CK+.

TABLE IV
RESULTS OF SRNN, TRNN, AND STRNN ON SEED

Model	number of channels	number of frequency bands	accuracy (%)
SRNN	62	5	85.88/9.98
TRNN	62	5	85.20/9.13
STRNN	62	5	**89.50/7.63**

TABLE V
RESULTS OF STRNN AND NONSPARSE STRNN

Dataset	Method	Accuracy
SEED	non-sparse STRNN	88.1
	STRNN	**89.5**
CK+	non-sparse STRNN	94.2
	STRNN	**95.4**

Fig. 6. Example of the weight distribution over facial regions.

3DCNN-DAP can achieve 92.4% which benefits from using strong spatial structural constraints on the dynamic action parts. DTAN employed in [37] achieves the accuracy of 91.4% by applying a CNN model which is able to capture temporal changes of appearance. Our STRNN achieves 95.4% which is more competitive to these state-of-the-art methods following tenfold cross validation protocol.

Fig. 5 shows the confusion matrix for CK+. In general, our algorithm performs well in recognizing all types of emotion as the accuracy of each expression is more than 90%. Among them, four kinds of expressions including anger, happiness, sadness, and surprise are relatively easy to be recognized with the accuracies of 97.73%, 97.06%, 96.00%, and

96.47%, which may attribute to their relatively large muscle deformations. Next ones are contempt, disgust, and fear, respectively, with 93.75%, 91.94%, and 91.67% recognition rates. Relatively high confusions appear between three pairs of expressions: 1) contempt versus angry; 2) contempt versus fear; and 3) contempt versus sadness, which may be intuitively due to the similar muscle deformations.

V. Conclusion

In this paper, a novel STRNN method is proposed to deal with EEG signal-based and face image-based human emotion recognition. To well model spatial co-occurrent variations and time dependent variation of human emotions, a multidirection SRNN layer and a bi-direction TRNN layer are hierarchically employed to learn spatial and temporal dependencies layer by layer. To adapt the multichannel EEG signals to the proposed STRNN framework, the spatial scanning order of electrodes are specified by spatial coordinates and temporal variation information is involved by slicing a window on the extracted DE feature sequences. To further select those salient regions of emotion representation as well as increase the model discriminant ability, we impose sparse projection onto those hidden states of spatial and temporal domains. The experimental results on both SEED EEG dataset and CK+ facial expression dataset have demonstrated that the proposed STRNN method achieves the state-of-the-art performance.

As the STRNN method can be seen as an integration of both SRNN and TRNN, it is still interesting to see how much improvement could be obtained by RNN modeling in the spatial or temporal domains or how much performance improvement can be gained by adding sparse constraints in STRNN. In addition, it is also interesting to see what the salient regions learned by the sparse STRNN would be located in the facial expression sequences. To answer all of these questions, we will also conduct additional experiments in what follows.

1) Comparisons of STRNN With SRNN and TRNN: To compare STRNN with SRNN and TRNN, the STRNN method is revised into only SRNN and only TRNN, in which process the remaining SRNN or TRNN is merged into a full connection layer. The results are shown in Table IV. The network which contains only TRNN achieves the accuracy of 85.20% with the deviation of 9.13%, while the network containing only SRNN achieves a little higher accuracy of 85.88% but with a higher deviation of 9.98%. STRNN achieves the accuracy of 89.50% which is about 4% higher than SRNN or TRNN with a lower deviation. The improvement of performance demonstrates the effectiveness of the hierarchical structure of SRNN and TRNN layers which learns both spatial and temporal dependencies.

2) Comparisons of STRNN With Nonsparse STRNN: To compare STRNN with nonsparse STRNN, we conduct experiments using the proposed STRNN method and an STRNN method without sparse constraints on projection matrices. The results are shown in Table V. As we can see, for both SEED and CK+ datasets, the accuracies of STRNN are about more than one percent higher than those of nonsparse STRNN, which verifies the effectiveness of sparse constraints which improve the performance of the proposed STRNN as well as achieve salient emotion regions detection.

3) Salient Emotion Detection: In addition to showing average recognition accuracies, we also visualize the weights of hidden states of the multidirection SRNN layer in our STRNN in the experiment conducted on CK+ dataset. In this process, the columns of the absolute coefficient values of projection matrices $\mathbf{G}^r$ are averaged over all spatial traversing directions. Fig. 6 shows the distribution of detected salient facial regions by mapping the weights of hidden states back to corresponding spatial regions in a 2-D facial image. As it is shown, the highlighted parts in the left image of Fig. 6 correspond to the action regions around mouth, eyes, and nose, which are intuitively crucial for human to perceive facial expression. Moreover, the black regions in Fig. 6 indicate that most values in projection matrices are near zero, which verifies the effectiveness of l_1-norm terms in the loss function for ensuring the sparsity of the column vectors of projection matrices.

References

[1] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[2] 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. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis.*, vol. 1, 2015, p. 6.

[5] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Comput.*, vol. 1, no. 2, pp. 270–280, 1989.

[6] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 186–197, Mar. 2010.

[7] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. IEEE Conf. Neural Eng.*, San Diego, CA, USA, 2013, pp. 81–84.

[8] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Disc.*, vol. 2, no. 2, pp. 121–167, 1998.

[9] Y. Li, W. Zheng, Z. Cui, and X. Zhou, "A novel graph regularized sparse linear discriminant analysis model for EEG emotion recognition," in *Proc. Int. Conf. Neural Inf. Process.*, Kyoto, Japan, 2016, pp. 175–182.

[10] W. Liu, W. L. Zheng, and B. L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.*, Kyoto, Japan, 2016, pp. 521–529.

[11] W. Zheng, "Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis," *IEEE Trans. Cogn. Develop. Syst.*, vol. 9, no. 3, pp. 281–290, Sep. 2017, doi: 10.1109/TCDS.2016.2587290.

[12] Y. Cai *et al.*, "Video based emotion recognition using CNN and BRNN," in *Proc. Chin. Conf. Pattern Recognit.*, 2016, pp. 679–691.

[13] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, Leeds, U.K., 2008, pp. 1–10.

[14] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.

[15] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Learning bases of activity for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1965–1978, Apr. 2017.

[16] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Win. Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[17] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. ACM Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 357–360.

[18] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1749–1756.
[19] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 143–157.
[20] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 1764–1772.
[21] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTERSPEECH*, vol. 2, 2010, pp. 1045–1048.
[22] G. Keren and B. Schuller, "Convolutional RNN: An enhanced model for extracting features from sequential data," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Vancouver, BC, Canada, 2016, pp. 3412–3419.
[23] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 2392–2396.
[24] S. O. Arik *et al.* (2017). *Convolutional Recurrent Neural Networks for Small-Footprint Keyword Spotting*. [Online]. Available: https://arxiv.org/abs/1703.05390
[25] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1110–1118.
[26] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Vancouver, BC, Canada, 2013, pp. 6645–6649.
[27] Z. Zuo *et al.*, "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Boston, MA, USA, 2015, pp. 18–26.
[28] F. Visin *et al.* (2015). *ReSeg: A Recurrent Neural Network for Object Segmentation*. [Online]. Available: https://arxiv.org/abs/1511.07053v1
[29] F. Visin *et al.* (2015). *ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks*. [Online]. Available: https://arxiv.org/abs/1505.00393
[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[31] P. Lucey *et al.*, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, 2010, pp. 94–101.
[32] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2012.
[33] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[34] R. Ptucha, G. Tsagkatakis, and A. Savakis, "Manifold based sparse representation for robust expression recognition without neutral subtraction," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Barcelona, Spain, 2011, pp. 2136–2143.
[35] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Tampa, FL, USA, 2013, pp. 103–110.
[36] S. Jain, C. Hu, and J. K. Aggarwal, "Facial expression recognition with temporal modeling of shapes," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Barcelona, Spain, 2011, pp. 1642–1649.
[37] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 2983–2991.

Tong Zhang received the B.S. degree from the Department of Information Science and Technology, Southeast University, Nanjing, China, in 2011, the M.S. degree from the Research Center for Learning Science, Southeast University, Nanjing, in 2014, where he is currently pursuing the Ph.D. degree in information and communication engineering.

His current research interests include pattern recognition, machine learning, and computer vision.

Wenming Zheng (M'08) received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004.

Since 2004, he has been with the Research Center for Learning Science, Southeast University, where he is currently a Professor with the Key Laboratory of Child Development and Learning Science, Ministry of Education, Research Center for Learning Science. His current research interests include neural computation, pattern recognition, machine learning, and computer vision.

Zhen Cui received the B.S. degree from Shandong Normal University, Jinan, China, in 2004, the M.S. degree from Sun Yat-sen University, Guangzhou, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014.

He was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2014 to 2015. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. His current research interests include sparse coding, manifold learning, deep learning, face detection, alignment and recognition, and image super resolution.

Yuan Zong received the B.S. and M.S. degrees in electronics engineering from Nanjing Normal University, Nanjing, China, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Research Center for Learning Science, Southeast University, Nanjing.

His current research interests include affective computing, pattern recognition, and speech signal processing.

Yang Li received the B.S. degree in electronic information and science technology from the School of Physics and Electronics, Shandong Normal University, Jinan, China, in 2012, the M.S. degree in electronic and communication engineering from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2015. He is currently pursuing the Ph.D. degree in information and communication engineering, Southeast University, Nanjing, China.

His current research interests include pattern recognition and machine learning.