Sensors and Materials, Vol. 32, No. 4 (2020) 1261–1277
MYU Tokyo

S & M 2175

Unsupervised Recurrent Neural Network with Parametric Bias Framework for Human Emotion Recognition with Multimodal Sensor Data Fusion

Jie Li,[1] Junpei Zhong,[2] and Min Wang[1*]

[1]School of Automation Science and Engineering, South China University of Technology,
381 Wushan Road, Tianhe District, Guangzhou, 510641, China
[2]School of Science and Technology, Nottingham Trent University,
Nottingham, NG11 8NS, UK

(Received August 9, 2019; accepted October 16, 2019)

Keywords: emotion recognition, multimodal sensors, recurrent neural network, subconscious behaviors

In this paper, we present an emotion recognition framework based on a recurrent neural network with parametric bias (RNNPB) to classify six basic emotions of humans (joy, pride, fear, anger, sadness, and neutral). To capture the expression to recognize emotions, human joint coordinates, angles, and angular velocities are fused in the process of signal preprocessing. A wearable Myo armband and a Kinect sensor are used to collect human joint angular velocities and angles, respectively. Thus, a combined structure of various modalities of subconscious behaviors is presented to improve the classification performance of RNNPB. To this end, two comparative experiments were performed to demonstrate that the performance with the fused data outperforms that of the single modality sensor data from one person. To investigate the robustness of the proposed framework, we further carried out another experiment with the fused data from several people. Six types of emotions can be basically classified using the RNNPB framework according to the recognition results. These experimental results verified the effectiveness of our proposed framework.

1. Introduction

Emotions have an important effect on a person's daily life. It is crucial to read emotions accurately and effectively from other people to avoid misunderstanding in interpersonal interactions. The ability of perceiving, understanding, and handling of one's own and others' emotions can be regarded as an expression of emotional intelligence.(1) It is one of the important abilities for individual survival. Moreover, available studies have shown that the skills of emotional intelligence have a high correlation with our mental health.(2) Emotion recognition using computing techniques has attracted increasing attention in recent years. The applications of emotion recognition, such as human–robot interaction (HRI), autonomous driving vehicles, intelligent surveillance systems, and entertainment, are very popular in our lives.(3–8) For an intelligent robot, endowing it with the ability of emotion recognition and cognition is very

*Corresponding author: e-mail: auwangmin@scut.edu.cn
https://doi.org/10.18494/SAM.2020.2552

ISSN 0914-4935

helpful for detecting and identifying human emotional states, reasoning, making decisions, and reacting to human expressions appropriately in HRI. For example, the capability of understanding unspoken intentions or feelings exactly through autistic children's physical behavior can help a robot grasp their mental status and adjust the topic timely as needed in the interactive communication process.[9]

As a complicated mental state, emotions often result in physical and psychological changes. These changes are associated with many internal and external activities. The internal activities include electroencephalogram (EEG), electrocardiograph (ECG), and electromyography (EMG) signals. The external activities involve body languages that are affected, mediated, and even regulated by emotions. In fact, body language, especially sensorimotor behavior, is usually subconscious; thus, it is rarely deceptive. Therefore, sensorimotor behaviors can be used to distinguish different emotions.

Various types of features have been utilized to recognize emotion successfully by different modeling methods for these features. These features include human facial expression, text, voice intonation, and some physiological signals, such as EEG and electrooculography (EOG). From these cues, one of the most popular features is facial expression. A number of emotion classification methods based on facial expression have been studied.[10–13] The commonality of these methods is that the features usually are appearance features, geometric features, or a hybrid of appearance and geometric features of the target face. For the appearance features, the information that describes the texture of the face is often extracted from different face or global face regions.[10,11] The geometric features are usually constructed as a feature vector by using the relationship between different facial components.[12] As for the hybrid features, the authors combined the advantages of appearance and geometric features to provide better results in certain cases.[12,13]

Although numerous studies mainly focus on facial expressions, there is increasing attention on other channels such as EEG, voice, and text.[14,15] Some advanced approaches have also been explored and developed to prove that multimodal information outperforms a single modality in recognition results.[16] However, most of the previous studies concentrated on supervised methods to recognize emotion using labeled datasets, and few studies focused on unsupervised methods by using human behaviors from ordinary users. As with any supervised learning problem, once we pick a model to classify emotions, it is difficult to obtain a labeled and sufficiently large training set. First, collecting emotion data and tagging those huge data are very troublesome and time-consuming. Second, we have to take someone's true emotion into account to evaluate the effectiveness of the data. Since videos or other signals do not always generate corresponding emotions for the user, nobody is sure whether the features are sufficiently reliable before feeding into algorithms. Aside from that, there is still another problem, that is, the interface of the device is often unfriendly and inconvenient to acquire data. To address the problem of tagging huge data, an unsupervised method with generalization ability is a promising solution. The wearable devices that are easy to use for ordinary users provide an alternative way to collect and train the data in our daily life. Hence, a possible solution to address these problems is using the unsupervised method to recognize emotion with more believable emotion features collected by a wearable device.

The wearable devices are usually used to capture sensorimotor behaviors, particularly human joint movements. Human joints from sensorimotor behaviors have been reported to be one of the critical features of emotion.[17] In our previous study, we applied the continuous joint coordinates from human nonverbal behavior to classify five emotions (joy, pride, fear, anger, and sadness) using unsupervised methods.[18] However, we captured behavior using only the Kinect sensor, which does not consider the advantages of the wearable device and multimodal data fusion. Many studies have shown that multimodal information can improve recognition performance.[19] It is also an interesting challenge to merge different modalities of information together and apply data fusion technologies to achieve the purpose of understanding emotions. On the other hand, most researchers have concentrated on integrating auditory and visual modalities to recognize emotion.[20] In contrast, a few research efforts have centered around human joints in multimodal emotion recognition, such as human joint angles and joint angular velocity. Compared with physiological signals and videos, different modalities of information of the joint convey more abundant and essential cues of human emotional states.

In this paper, to integrate the spatial and continuous temporal features of human joints, we present an unsupervised framework called the recurrent neural network (RNN) with parametric bias (RNNPB) to perceive six emotions (joy, pride, fear, anger, sadness, and neutral). The Kinect sensor was used to obtain joint coordinates and angles, and the Myo armband was used to collect the joint angular velocity. The main contributions of this paper are summarized as follows.

(1) Compared with other emotion recognition methods, multimodality signals using Kinect and Myo armbands are employed to achieve an easy and fast deployment of the sensors. We also demonstrate that using these two sensors leads to more accurate results in our learning framework.
(2) Human emotions are recognized by bodily behaviors using an unsupervised framework. This framework can overcome the disadvantages of usual supervised emotion recognition methods that need a large number of labeled training data.
(3) Because of the generalization ability of the proposed framework, six untrained emotional behaviors (joy, pride, fear, anger, sadness, and neutral) collected from different people are well classified.

2. Preliminary

In this section, we first introduce the framework of the proposed method. Some relevant devices for acquiring data are also described in detail.

2.1 General overview

The framework of emotion recognition by RNNPB is presented in Fig. 1. A Kinect sensor and a Myo armband were used to capture different modalities of human behaviors. Joint coordinates, angles, and angular velocities of human behaviors were simultaneously collected while people were presenting certain actions in the process of data collection.

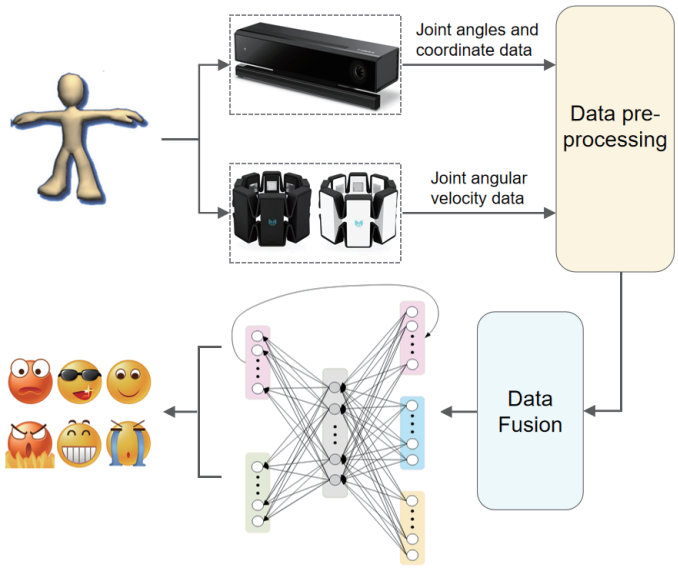


Fig. 1. (Color online) Framework of our proposed method.

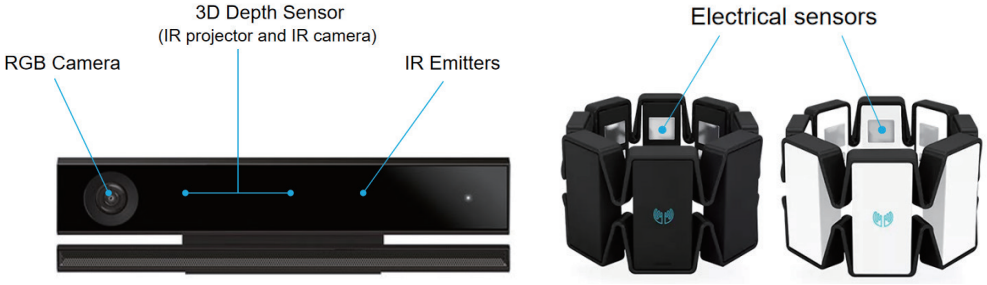


Fig. 2. (Color online) (a) Kinect V2 sensor and (b) Myo armband.

2.2 Kinect sensor

The Kinect for Windows v2 sensor (Kinect V2) was used in our work. It contains three vital pieces: an RGB color camera, an IR emitter, and a 3D depth sensor to provide color, IR, and depth images, as shown in Fig. 2(a). With these devices, the Kinect sensor can track up to human skeletons, capture full-body 3D motion, and recognize simple gestures. Compared with Kinect V1, Kinect V2 can track 25 body joints. In this paper, Kinect V2 was used to collect 3D joint coordinates and angles of human behaviors.

2.3 Wearable device (Myo armband)

The Myo armband [Fig. 2(b)] is a body-wearable and portable device produced by Thalmic Labs. It is a lightweight elastic armband consisting of a number of metal contacts. These metal contacts can measure electrical activity in a user's forearm muscle to transmit gestures that he/she makes with his/her hands to a control computer via Bluetooth. Therefore, the Myo armband

allows the user to control his/her cell phones, computers, and other favorite digital technologies wirelessly with hand gestures and motions by reading the electrical activity of muscles and the motion of the arm. Hand gestures and motions are detected by proprietary EMG muscle sensors and a highly sensitive motion sensor separately. The Myo armband is used to capture the joint angular velocity of the human arm.

3. Data Collection

In this section, the specific process of data collection will be described; this process includes joint coordinates, human upper body joint angles, and joint angular velocities from emotion-aroused human body behavior. Since the Kinect sensor can capture human joint coordinates directly, the details on how to acquire the joint coordinates of a human body will not be introduced.

3.1 Joint angles captured by Kinect sensor

The Kinect sensor can track up to six people's whole skeletons within its view at one time. Each skeleton has 25 joints. These joints are numbered 0–24 [Fig. 3(a)]. Through the RGB camera and depth sensor of Kinect, we can acquire the 3D coordinates of each joint for an object human body. As shown in Fig. 3(b), skeletons can be tracked regardless of whether the object human body is standing or sitting. Note that the Kinect sensor treats joints as one person is looking in the mirror. Thus, the "left side" human body joints are on the left in Fig. 3 and the "right side" human body joints are on the right.

Once we obtain the 3D coordinates of the human joints, the joint angles can be calculated by the space vector approach. Assuming that there are two points $P=(x_0, y_0, z_0)$ and

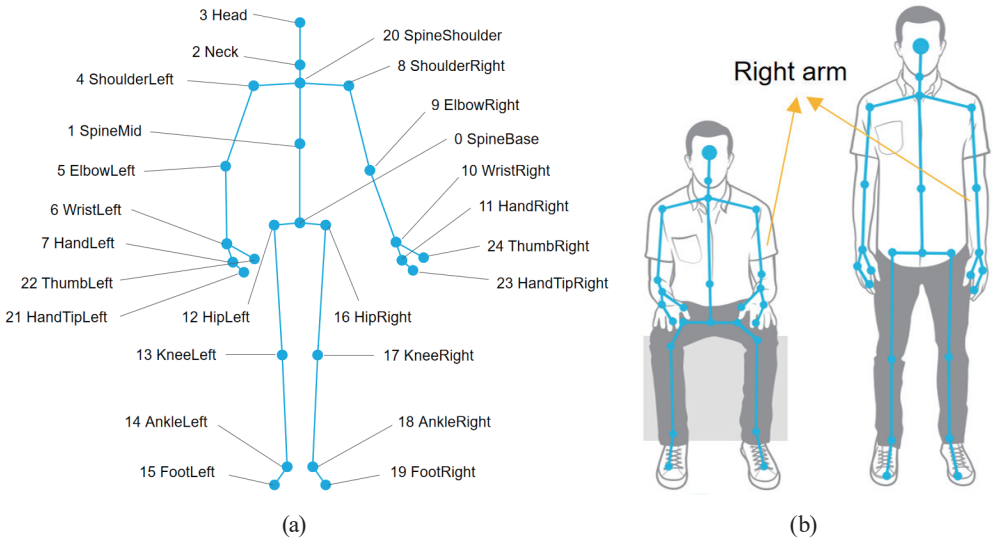


Fig. 3. (Color online) (a) 25 human joints and (b) human body skeleton captured by Kinect sensor.

$Q = (x_1, y_1, z_1)$ in 3D space, the distance between these two points can be calculated as

$$d_{PQ} = \left|\overrightarrow{PQ}\right| = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}, \tag{1}$$

where vector $\overrightarrow{PQ} = (x_1 - x_0, y_1 - y_0, z_1 - z_0)$ and d_{PQ} is the distance between the points P and Q. Using the law of cosines, the angle between two vectors can be calculated easily.[21] Similarly, the angle between two joints can be obtained by applying the same method. In the Kinect coordination, a joint can be regarded as a vector. Assume that joint 1 is expressed as $\overrightarrow{OA}$ and joint 2 is expressed as $\overrightarrow{OB}$; then, the angle between these two joints can be computed as

$$\cos\angle \mathrm{AOB} = \cos(\overrightarrow{OA}, \overrightarrow{OB}) = \frac{\overrightarrow{OA} \cdot \overrightarrow{OB}}{\left|\overrightarrow{OA}\right| \cdot \left|\overrightarrow{OB}\right|}. \tag{2}$$

According to Eq. (1), the coordinates obtained by the Kinect sensor can be converted to vectors and the corresponding angles can be calculated using Eq. (2).

In this work, only the upper human body joint angles consisting of left and right arms joint angles were collected. Since each arm has seven degrees of freedoms (DoFs), 14 joint angles were captured in total, which include the shoulder pitch angle, shoulder roll angle, shoulder yaw angle, elbow pitch angle, elbow roll angle, wrist pitch angle, and wrist yaw angle for both left and right arms. Figure 4 shows the specific angle calculation process of a left arm based on the space vector approach. The black dotted lines OX, OY, and OZ are the Kinect's 3D coordinate system in Cartesian space, and the red dotted lines are auxiliary lines. The shoulder pitch angle $\angle$COD was computed using Eq. (2) from the vector $\overrightarrow{CO}$ to $\overrightarrow{CD}$, and the points C, O, and D denote the left shoulder, left hip, and left elbow, respectively. The same method was utilized

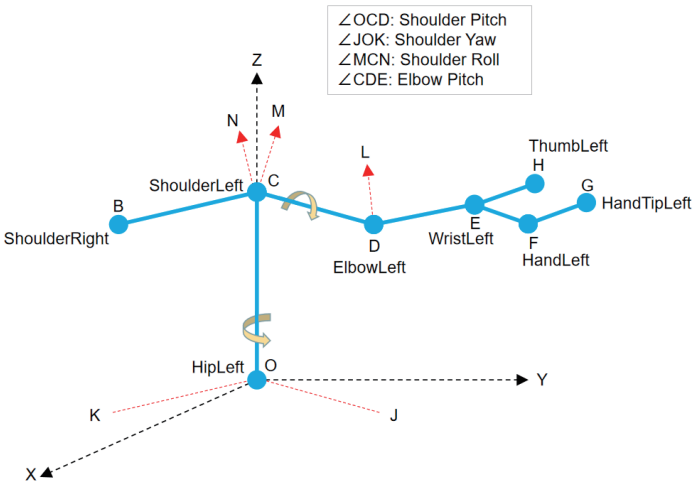


Fig. 4. (Color online) Human joint angles of left arm calculation.

to calculate the elbow pitch angle ∠CDE and wrist pitch. Similarly to the shoulder pitch, the shoulder yaw angle ∠JOK can be computed by applying the three points B, C, and D, which represent the right shoulder, left shoulder, and left elbow, respectively. In Fig. 4, the vectors $\overrightarrow{OK}$ and $\overrightarrow{OJ}$ with the red dotted line are the mapping results of the vectors $\overrightarrow{CB}$ and $\overrightarrow{CD}$ in the plane *XZ*, and are separately parallel to $\overrightarrow{CB}$ and $\overrightarrow{CD}$. Therefore, the wrist yaw angle can be calculated by the same method. The elbow roll angle can be obtained by computing the angle of two planes between CDE and DEF by the same method. Among the auxiliary lines, $\overrightarrow{DL}$ and $\overrightarrow{CM}$ are respectively vertical to the vectors $\overrightarrow{DE}$ and $\overrightarrow{CD}$, and the vector $\overrightarrow{CN}$ is parallel to $\overrightarrow{DL}$. Then, the angle ∠MCN between the vectors $\overrightarrow{CN}$ and $\overrightarrow{CM}$ is defined as the shoulder roll, which can be calculated by a similar calculation method.

The corresponding angles of the right arm were computed in the same way. Thus, the angles of human body behaviors were acquired and these angles were fed into the unsupervised algorithm together with other modality data to perceive human emotions.

3.2 Joint angular velocity collection by Myo armband

To obtain the joint angular velocity, human subjects need to wear two Myo armbands for each arm. One of the Myo armbands is worn near the center of the forearm and the other one is worn near the center of the upper arm. The Myo armband uses quaternions to obtain the joint angle and then collects the joint angular velocity by computing the basic change in joint angle. According to Yang *et al*., if the relevant joint angles are zero, any position of the human arm can be regarded as the initial position.[21] When the human arm is moved to a new position *U*, the corresponding angle from the initial position to pose *U* is the rotation angle, namely, the joint angle.

We assume that the initial orientation of the Myo armband is denoted by frame (X_{l1}, Y_{l1}, Z_{l1}), and that the current orientation of the Myo armband is denoted by frame (X_{l2}, Y_{l2}, Z_{l2}). Then, the angular velocities of the shoulder pitch v_{lx}, shoulder roll v_{ly}, and shoulder yaw v_{lz} can be obtained by the Myo armband worn on the left upper arm. The angular velocities of the elbow pitch v_{l2x} and elbow roll v_{l2y} were acquired by the Myo armband worn on the left forearm. Then, five joint angular velocities of the right arm, v_{lx}, v_{ly}, v_{lz}, v_{l2x}, and v_{l2y}, can be acquired in the same way. Ten joint angular velocities for human arms were taken from the Myo armband in total.

Before collecting data, the participants were required to stand in front of the Kinect sensor wearing two Myo armbands for each arm. After that, the training data was captured by the devices with two computers while the participants were showing emotional behaviors. One was used to obtain the joint coordinates and angles; the other was used to obtain the joint angular velocities of the left and right arms separately. For each emotion, two sequences were collected from one person. The data collection experiments included four healthy participants aged between 22 and 30 years (two females and two males). The participants were asked to perform six types of emotion-aroused behaviors in our experiments. There were 48 sequences from four participants in total.

4. Methods

4.1 Preprocessing

Each joint data obtained from the Kinect sensor has eleven properties: color coordinates (X, Y), depth coordinates (X, Y), camera coordinates (X, Y, Z), and orientation coordinates (X, Y, Z, W). The Kinect's camera coordinates use the Kinect's infrared sensor to find the 3D points of the joints in space, and the camera space refers to the 3D coordinate system used by the Kinect. In this paper, we focused on the camera coordinates, which are needed to obtain 3D coordinate data. Nine joint coordinates (head, neck, torso, right shoulder, left shoulder, right elbow, left elbow, right hand, and left hand) from the human upper body were collected since they are significant for emotion. In other words, the dimensions of joint coordinates were 27.

As mentioned above, there are 24 features from human arms, which include 14 joint angles and 10 joint angular velocities. For modality fusion, the feature-level fusion was employed to concatenate three types of feature vectors into a larger feature vector. The total number of dimensions of emotion-aroused human behaviors was 51.

4.2 Unsupervised emotion recognition methods

RNNPB, as an unsupervised learning method, was employed to learn multimodal sensorimotor behaviors and classify human emotions by the corresponding spatiotemporal sequences.[22,23] RNNPB is substantially a RNN of the Jordan or Elman type. Here, the Elman-type RNN architecture was used.[24]

Figure 5 shows the structure of unsupervised RNNPB of the Elman type.[23] This RNNPB consists of five types of layers: input layer, hidden layer, parametric bias units (PB layer), context layer, and output layer. The input of hidden layers (y_h) includes three parts; the details are expressed as

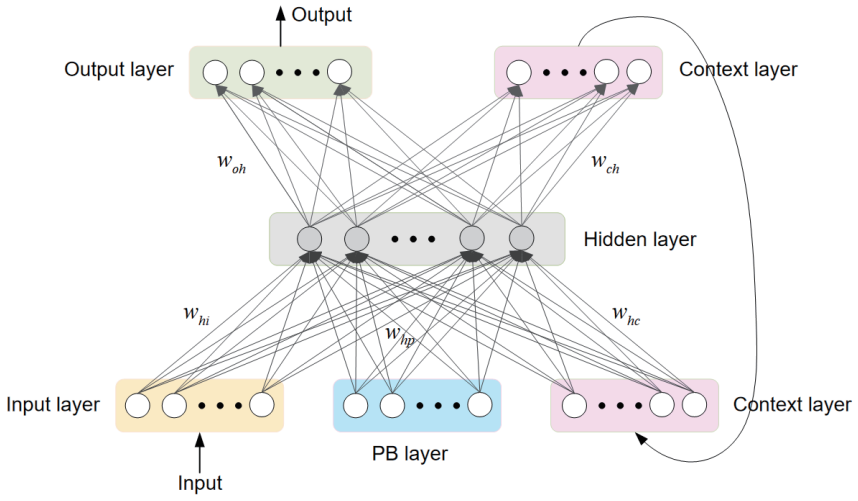


Fig. 5. (Color online) Structure of RNNPB in this paper.

$$y_h(k) = \sum_i g_i(k) w_{hi} + \sum_c g_h(k-1) w_{hc} + \sum_n PB_n(k) w_{hp} , \tag{3}$$

where k is the time step, and k is omitted to avoid agitation if the parameters express the states from the same time step in one expression. w_{hi} is the weight between the hidden layer and the input layer, w_{hp} is the weight between the PB layer and the hidden layer, and the weight connecting the hidden layer with the context layer is denoted as w_{hc}. $g_i(k)$ is the activation function, and the subscripts i and h are related to the parameters of the input and hidden layers. $PB_n(k)$ depicts the activation function of the PB layer.

Figure 5 clearly shows that the PB layer is connected with hidden layers. The internal values of the PB layer are adjustable. Also, the PB layer endows the network with generalization ability. Most importantly, the trained values of the PB layer could be applied to recognize multiple sequences such as emotional behaviors. The internal values of the additional PB layer are learned unsupervised and updated through back-propagation in a self-organized manner. Also, the weights of this network are updated with backpropagation through time (BPTT). The internal values of the PB layer at the time step k of the i-th time series are updated as

$$\rho_i(e+1) = \rho_i(e) + \gamma_i \sum_{k=1}^{T} \delta_{i,k}^{PB} , \tag{4}$$

where the epoch e represents an entire forward-backward training cycle, $\rho_i(e)$ is the output of PB layers, $\delta_{i,k}^{PB}$ is the backpropagation error of the PB layer at the time step k, and T is the length of each time series. γ_i is the update rate of the PB layer, and the relationship between γ_i and $\delta_{i,k}^{PB}$ is formalized as

$$\gamma_i = \frac{1}{T} \left\| \sum_{k=1}^{T} \delta_{i,k}^{PB} \right\| . \tag{5}$$

The cost function during training is determined by

$$J = \frac{1}{2} \sum_k^T \sum_i^N (g_i^d(k+1) - g_i^o(k))^2 , \tag{6}$$

where $g_i^d(k+1)$ is the desired output, $g_i^o(k)$ is the actual output, and N is the size of the output layer. The weights in the network obey the gradient descent and will be updated by the following equation:

$$\Delta w_{ij} = -\gamma_{ij}(e) \frac{\partial J}{\partial w_{ij}} . \tag{7}$$

The learning rate of weights (γ_{ij}) is adjusted using the partial derivative of w_{ij} after every epoch. The partial derivative of w_{ij} can be positive or negative, which means that the sign is

changing. The change in sign is determined by

$$\varepsilon_{ij} = \frac{\partial J}{\partial w_{ij}}(e-1)\frac{\partial J}{\partial w_{ij}}(e). \tag{8}$$

If $\varepsilon_{ij} > 0$, the learning rate has to be increased by a factor, which is greater than one, to speed up convergence, and vice versa. The update of the learning rate can be expressed as

$$\gamma_{ij}(e) = \begin{cases} \max(\gamma_{ij}(e-1)\cdot\zeta^{-}, \gamma_{min}) & \text{if} \quad \varepsilon_{ij} > 0, \\ \min(\gamma_{ij}(e-1)\cdot\zeta^{+}, \gamma_{max}) & \text{if} \quad \varepsilon_{ij} < 0, \\ \gamma_{ij}(e-1) & \text{else.} \end{cases} \tag{9}$$

Here, ζ^{-} and ζ^{+} represent the changing rate of γ_{ij}, and $\zeta^{-} < 1$ is the decreasing rate, $\zeta^{+} > 1$ is the increasing rate, and γ_{min} and γ_{max} are the minimum and maximum values of γ_{ij}, respectively.

The sigmoid function proposed in Ref. 25 is used for all neurons in RNNPB, as well as for the transfer function in the PB layer:

$$\text{sigmoid}(x) = a\cdot\tanh\left(\frac{2}{3}x\right),\ a = 1.7159, \tag{10}$$

$$PB_k = \text{sigmoid}\left(\frac{2}{3}\rho_k(e)\right), \tag{11}$$

where x denotes the input vector to the neurons in the hidden and output layers.

The RNNPB model is used to classify human emotions without the labeled datasets. For this method, the values of PB units indicate the corresponding emotions of datasets. Different sequences with the same emotion will result in similar PB values based on the method. Thus, human emotions are recognized in an unsupervised way. Because of the additional PB layer, the RNNPB model is endowed with generalization ability to untrained datasets. This means that although few samples are trained, a relatively stable recognition result will be obtained. Before the fused data is fed into the network, normalization is needed for the input features to enhance the accuracy and convergence speed of the model. The values of the normalized datasets range from zero to one. The multimodal RNNPB model is implemented in Python language.

5. Experiments

Motivated by our previous work,[17] we performed three experiments to recognize six human emotions and compare the clustering performance in different cases. The details will be introduced as follows.

5.1 Experimental setup

The experiments were performed with the same parameters for RNNPB to learn the spatiotemporal sequences of human behaviors. The parameters are shown in Table 1.

Except for the above parameters, the sizes of the input and output layers are not listed. The sizes of these two parameters are both equal to the dimensions of input data. Since the dimensions of the input data for each experiment are different, the sizes of the input and output layers are different.

5.2 Experimental results

Three experiments were implemented to explore how the different modality sensor data affect emotion recognition results. Three types of data sets were fed into RNNPB for training. In the first experiment, 12 sequences with 41 dimensions (two sequences for each emotion) that include the joint coordinates and angles were provided as the input of the network. For the second experiment, the same type of emotion data was used to recognize emotion with the same parameters for the network. Different from the first experiment, the dimensions of the input data were 51 and the additional 10 dimensions are human joint angular velocities including those of both the left and right arms. Note that all the training data sets regarding the first and second experiments were captured from one person. With respect to the third experiment, 24 sequences (four sequences for each emotion) expressing six emotions were trained to classify emotions. The data structure was the same as in the second experiment. However, the data sets were collected from four different people. Since the previous experiment was conducted using the single modal data (coordinates) based on RNNPB, only the results between the merging of information (joint coordinate and angle) and the fusion of different multimodal sensor data (data collected from the Kinect sensor and Myo armband) were compared in this paper.[16,17]

The PB values of the first and second experiments are shown in Figs. 6 and 7, respectively, and the corresponding results of the third experiment are presented in Fig. 8. The same shapes of the markers express the same motion, and the markers with different colors and the same shapes imply different sequences for one emotion in Figs. 6–8. The annotations of “angry1”, “angry2”, “angry3”, and “angry4” express different sequences for angry emotions in Figs.

Table 1
RNNPB parameters.

Parameter	Description	Value
H	Size of hidden layer	40
C	Size of context layer	40
P	Size of PB layer	2
γ_n	Learning rate of BPTT	0.001
γ_i	Learning rate of PB layer	0.2
ζ^-	Decreasing learning rate of γ_n	0.99999
ζ^+	Increasing learning rate of γ_n	1.00001
γ_{min}	Upper bound of γ_k	1×10^{-4}
γ_{max}	Lower bound of γ_k	1×10^{-8}

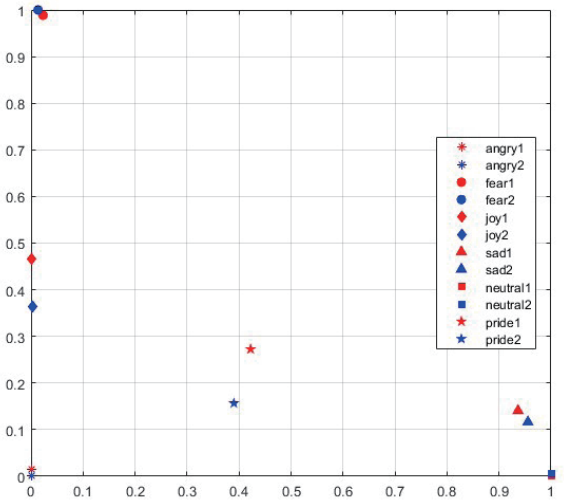


Fig. 6. (Color online) PB values of the first experiment in PB space.

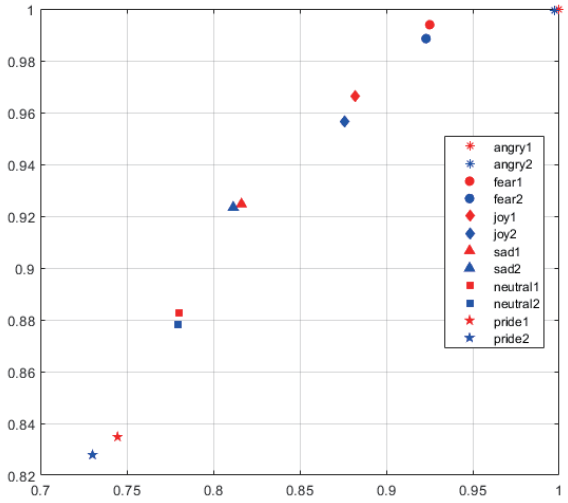


Fig. 7. (Color online) PB values of the second experiment in PB space.

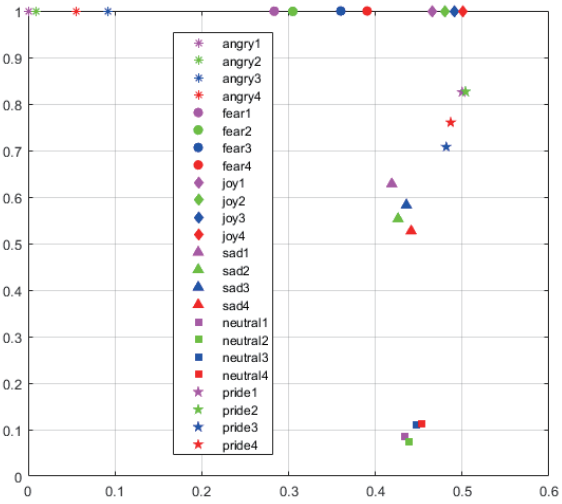
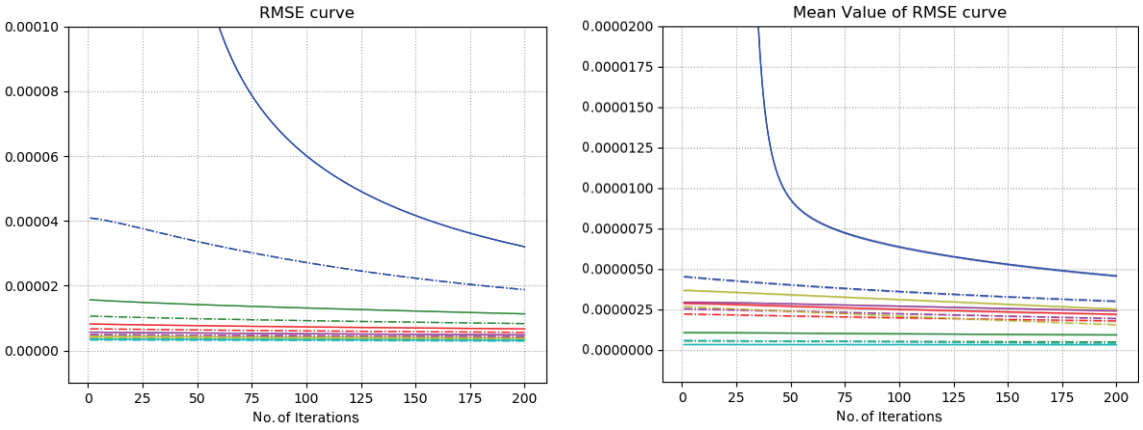


Fig. 8. (Color online) PB values of the third experiment in PB space.

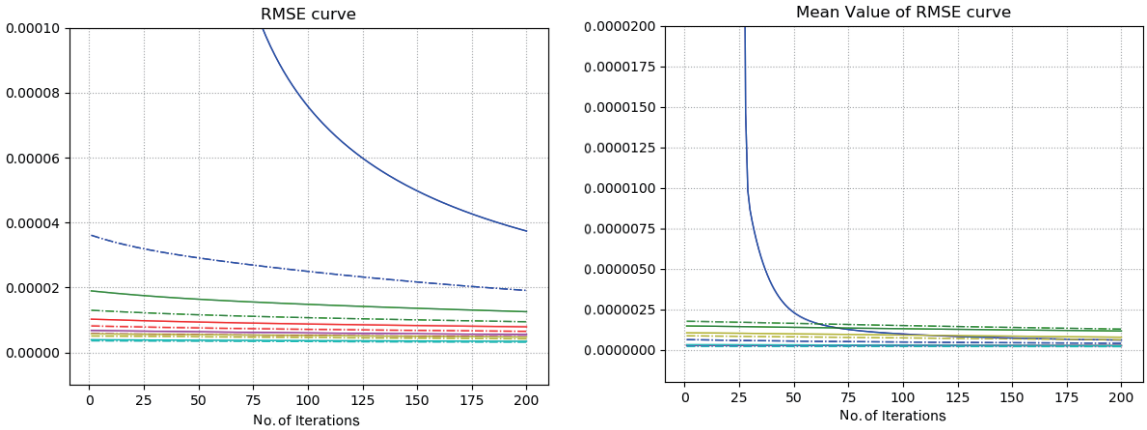
6–8. The annotations of the other four emotions are the same as those of angry emotions. Figures 9–11 show the root–mean–square error (RMSE) curves of the training process in 200 consecutive epochs (200 iterations for each epoch). For the RMSE curves, the same emotion is depicted by the same color with different shapes. Figures 9(a)–11(a) show the RMSE curve of the first epoch, and Figs. 9(b)–11(b) show the mean RMSE values of 200 epochs. In Figs. 9–11, the RMSE values are computed using the following equation:

$$RMSE(x, h) = \sqrt{\frac{1}{m}\sum_{i}^{m}(h(x^{(i)}) - y(i))^2}\ , \tag{12}$$



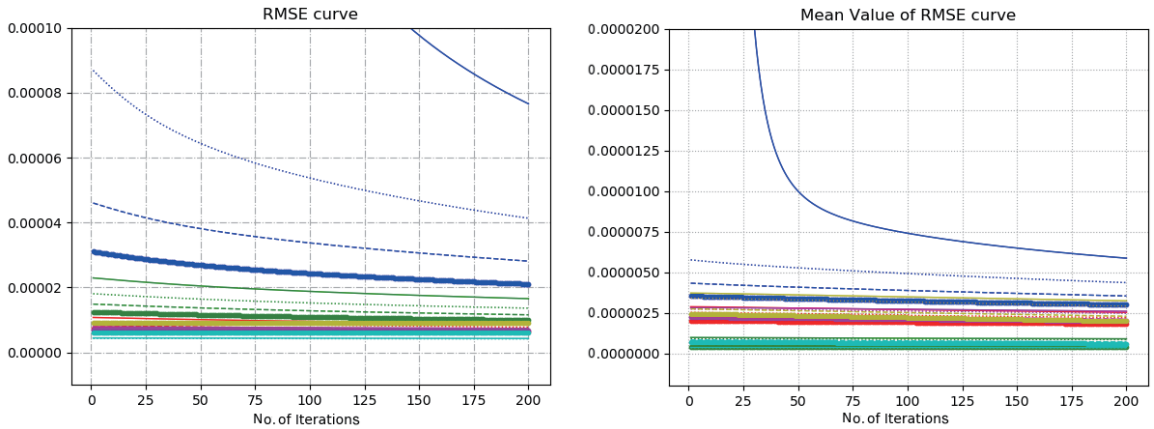
(a) (b)

Fig. 9. (Color online) (a) RMSE curve of epoch one and (b) mean values of RMSE curve of 200 epochs in the first experiment.



(a) (b)

Fig. 10. (Color online) (a) RMSE curve of epoch one and (b) mean values of RMSE curve of 200 epochs in the second experiment.



(a) (b)

Fig. 11. (Color online) (a) RMSE curve of epoch one and (b) mean values of RMSE curve of 200 epochs in the third experiment.

where m is the total number of samples, $h(x^{(i)})$ is the predicted value of the i-th sample, and $y(i)$ is the actual value of the i-th sample.

5.3 Analysis of experimental results

According to the experimental results, it is not difficult to find that the PB values corresponding to the same emotions are clustered together, and the RMSE is convergent to a small certain value. To investigate how the results vary with different modality data, the recognition performance characteristics of the first and second experiments were evaluated on the basis of the above results.

The emotion recognition performance was assessed from two perspectives. The first is the distance of PB values corresponding to different emotions. The quantitative confusion matrices are given in Figs. 12 and 13, which present the distance among various PB values in the PB space to evaluate the clustering results of the first and second experiments. Since the PB value in the PB space is a point, the distance between two PB values can be computed using Eq. (1). The distance of the PB values includes the intraclass distance d_w, interclass distance d_b, and relative distance d_r. d_r is calculated using the maximum d_b divided by the average of d_w. The intraclass distance reflects the aggregation level of the same class, and the interclass distance reveals the scattered level of different classes. The relative distance expresses the relationship between the intraclass and interclass distances. These distances reflect the clustering performance to some extent. In general, the desired clustering result is that d_w is small, and d_b and d_r are large. The second point is the convergence speed and eventual values of RMSE. The detailed analyses and comparisons will be discussed with respect to these two points.

Firstly, the specific distances of PB values are listed in Table 2 by observing Figs. 12 and 13. For the first experiment, it is clear that there is a large interclass distance among different

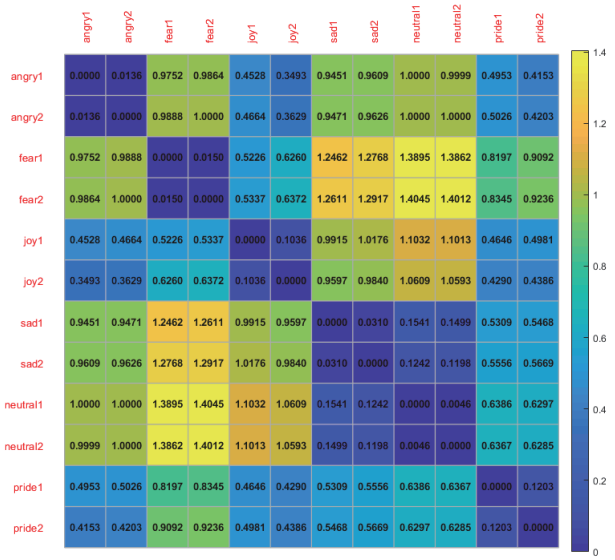


Fig. 12. (Color online) Distance matrices between different emotions for the first experiment.

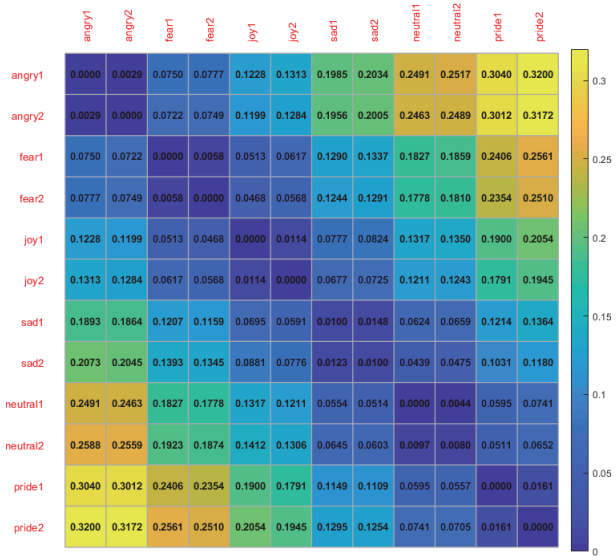


Fig. 13. (Color online) Distance matrices between different emotions for the second experiment.

Table 2
Distances of PB values in PB space for the first two experiments.

Description	First experiment	Second experiment
Average intraclass distance	0.048	0.0092
Largest interclass distance	1.4045	0.32
Relative distance between interclass and intraclass	29.26	34.78

emotions and a large intraclass distance from the same emotion (Fig. 6). As for the second experiment, the results in Fig. 7 clearly show that the intraclass and interclass distances are both smaller than those in the first experiment. However, the relative distance between the d_w and d_b of the second experiment is larger than that of the first one. Combining Fig. 7 and Table 2, we can conclude that the emotion recognition result of the second experiment is better. This implies that the joint angular velocity is useful for distinguishing different emotions and facilitating a smaller distance of the same class. In other words, the joint angular velocity may provide complementary information of emotions. Then, the convergence speed and the mean RMSE values of 200 epochs are compared by observing Figs. 9 and 10. We often expect a higher convergence speed and a smaller RMSE. In comparison with the first experiment, the convergence speed is much higher and the RMSE is slightly smaller than those in the second experiment. To sum up, the recognition results of the fused data from the multimodal sensor reveal a better performance than those of the single modal sensor data.

The first and second experiments merely classified six emotions from one person. Therefore, the third experiment was performed to investigate the effectiveness and stability of our proposed framework. Figure 8 shows the recognition results. The results imply that the RNNPB framework can basically classify six emotions from different people. Since there are differences in the behavioral expressions of different people to react to the same emotion, the classification results are slightly inferior to those of the first or second experiment. The expressions of sensorimotor behaviors are regulated by the internal emotion states;[16] there are also some common critical features for the same emotion of different people. This can be proved by the classification results. According to Figs. 6–8, we can find that the results of sad and neutral emotions are better than those of other emotions. That is probably because the external behavior and internal state of two emotions are both very similar.

6. Conclusions

In this paper, an unsupervised RNNPB framework was proposed to classify human emotions using multimodal sensor fused data. Multimodal data were the spatiotemporal sequences of emotional human behaviors, which were collected by a wearable Myo armband and a Kinect sensor containing human joint coordinates, angles, and angular velocities. Then, three experiments were performed to explore how multimodal data affect the emotion recognition results and to evaluate the stability of the RNNPB framework. The experimental results showed that multimodal fused data can markedly increase the relative distance between intraclass and interclass, and decrease the intraclass distance and RMSE compared with the single modal

sensor data. The qualitative and quantitative analysis and evaluation results demonstrated the effectiveness of our proposed RNNPB framework. Moreover, these experimental results also indicated that signals from different modalities provide complementary information, and that the multimodal information can be integrated to enhance the robustness of the emotion recognition system compared with a single modal framework.

In the future, we will combine visual information (facial expression), auditory information (voice), and human behaviors to construct a more robust and effective emotion recognition system to enhance the clustering performance.

Acknowledgments

This work was partially supported by the National Nature Science Foundation (NSFC) under Grants 61861136009 and 61811530281.

References

1 M. Ptaszynski, P. Dybala, W. H. Shi, R. Rzepka, and K. Araki: Proc. 2009 21st Int. Joint Conf. Artificial Intelligence (IJCAI, 2009) 1469. http://orcid.org/0000-0002-8274-0875
2 L. Zysberg: Psychology **9** (2018) 2471. https://doi.org/10.4236/psych.2018.911142
3 H. S. Koppula and A. Saxena: IEEE Trans. Pattern Anal. Mach. Intell. **38** (2016) 14. https://doi.org/10.1109/TPAMI.2015.2430335
4 C. Yang, X. Wang, L. Cheng, and H. Ma: IEEE Trans. Cybern. **47** (2017) 3148. https://doi.org/ 10.1109/TCYB.2016.2573837
5 M. S. Ryoo and J. K. Aggarwal: Int. J. Comput. Vision. **93** (2011) 183. https://doi.org/10.1007/s11263-010-0355-5
6 Y. Kong and Y. Fu: IEEE Trans. Pattern Anal. Mach. Intell. **38** (2016) 1844. https://doi.org/10.1109/TPAMI.2015.2491928
7 C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette: Speech Commun. **50** (2008) 487. https://doi.org/10.1016/j.specom.2008.03.012
8 Y. Kong, Z. Q. Tao, and Y. Fu: Proc. 2017 30th IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2017) 3662. https://doi.org/10.1109/CVPR.2017.390
9 K. M. Rump, J. L. Giovannelli, N. J. Minshew, and Strauss: Child Dev. **80** (2009) 1434. https://doi.org/10.1111/j.1467-8624.2009.01343.x
10 D9. Ghimire and J. Lee: Sensors **13** (2013) 7714. https://doi.org/10.3390/s130607714
11 S. L. Happy, A. George, and A. Routray: Proc. 2012 4th Int. Conf. Intelligent Human Computer Interaction (IEEE, 2012) 1. https://doi.org/10.1109/IHCI.2012.6481802
12 D. Ghimire, S. Jeong, J. Lee, and S. H. Park: Multimed Tools Appl. **76** (2017) 7803. https://doi.org/10.1007/s11042-016-3418-y
13 C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez: Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016) 5562. https://doi.org/10.1109/CVPR.2016.600
14 G. Peng, C. Yang, W. He, and C. L. P. Chen: IEEE Trans. Ind. Electron. **67** (2019) 1 (in press). https://doi.org/10.1109/tie.2019.2912781
15 A. Schirmer and R. Adolphs: Trends Cogn. Sci. **21** (2017) 216. https://doi.org/10.1016/j.tics.2017.01.001
16 W.-L. Zheng, W. Liu, Y. F. Lu, B.-L. Lu, and A. Cichocki: IEEE Trans. Cybern. **49** (2019) 1110. https://doi.org/10.1109/TCYB.2018.2797176
17 J. P. Zhong and L. Cañamero: Proc. 2014 4th Joint Int. Conf. Development and Learning and on Epigenetic Robotics (IEEE, 2014) 75. https://doi.org/10.1109/DEVLRN.2014.6982957
18 J. P. Zhong, R. Novianto, M. J. Dai, X. Z. Zhang, and A. Cangelosi: Proc. 2016 28th Chinese Control and Decision Conf. (IEEE, 2016) 4965. https://doi.org/10.1109/ccdc.2016.7531882
19 W.-L. Zheng, B.-N. Dong, and B.-L. Lu: Proc. 2014 36th Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society (IEEE, 2014) 5040. https://doi.org/10.1109/EMBC.2014.6944757
20 P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou: IEEE J. Sel. Top. Signal Process. **11** (2017) 1301. https://doi.org/10.1109/JSTSP.2017.2764438

21 C. Yang, G. Peng, L. Cheng, J. Na, and Z. Li: IEEE Trans. Syst. Man Cybern. Part A Syst. Humans (2019) 1. https://doi.org/10.1109/TSMC.2019.2920870 (in press)
22 T. Kuremoto, K. Morisaki, K. Kobayashi, S. Mabu, and M. Obayashi: Proc. 2014 2nd Int. Conf. Intelligent Systems and Image Processing (ICISIP, 2014) 414. https://doi.org/10.12792/icisip2014.078
23 J. Tani, M. Ito, and Y. Sugita: Neural Netw. **17** (2004) 1273. https://doi.org/10.1016/j.neunet.2004.05.007
24 J. Kleesiek, S. Badde, S. Wermter, and A. K. Engel: Communications in Computer and Information Science (Springer, Heidelberg, 2013) p. 83. https://doi.org/10.1007/978-3-642-36907-0_6
25 Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller: Lecture Notes in Computer Science (Springer, Heidelberg, 2012) p. 9. https://doi.org/10.1007/978-3-642-35289-8_3

About the Authors

Jie Li received her B.S. degree from Luoyang Normal University, Luoyang, China, in 2011 and her M.S. degree from Harbin Institute of Technology, Shenzhen, China, in 2014. She is currently pursuing her Ph.D. degree in control science in the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China. Her research interests are in affective computing, human–robot interaction, and machine learning. (jieLi_ac@163.com)

Junpei Zhong is currently an independent research fellow at Nottingham Trent University. Previously, he worked as a research scientist at the National Institute of Advanced Industrial Science and Technology (AIST), Waseda University, and Plymouth University. He received his B.S. degree from South China University of Technology in 2006, M.Phil from Hong Kong Polytechnic University in 2010, and Ph.D. ("with great distinction") from the University of Hamburg in 2015. His research interests include machine intelligence, cognitive robotics, and assistive robotics. (zhong@junpei.eu)

Min Wang received her B.Sc. degree in mathematics and M.Sc. degree in applied mathematics from Bohai University, Jinzhou, China, in 2003 and 2006, respectively, and her Ph.D. degree in system theory from Qingdao University, Qingdao, China, in 2009. She was a visiting scholar with the Department of Computer Science, Brunel University London, Uxbridge, United Kingdom, from 2017 to 2018. She is currently a professor at the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China. She has authored or coauthored over 40 papers published in international journals. She is currently an Associate Editor of IEEE Access, Control Theory & Applications. Her current research interests include intelligent control, dynamic learning, robot control, and event-triggered control. She is a very active reviewer for many international journals. (auwangmin@scut.edu.cn)