Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

CNN-based Broad Learning with Efficient Incremental Reconstruction Model for Facial Emotion Recognition ⋆

Luefeng Chen [*,**], **Min Li** [*,**], **Xuzhi Lai** [*,**,†], **Kaoru Hirota** [***], **Witold Pedrycz** [****]

[*] *School of Automation, China University of Geosciences, Wuhan 430074, China (e-mail: chenluefeng@cug.edu.cn)*
[**] *Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China (e-mail: laixz@cug.edu.cn)*
[***] *Tokyo Institute of Technology, Yokohama 226-8502, Japan (e-mail: hirota@jsps.org.cn)*
[****] *Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6R 2G7, Canada (e-mail: wpedrycz@ualberta.ca)*

Abstract: Convolutional neural network-based broad learning with efficient incremental reconstruction model (CNNBL) is proposed to recognize emotions in human-robot interaction. It aims to extract deep and abstract features from facial emotional images, and reduce the influence of the complex structure and slow network updates on facial emotion recognition in deep learning. Feature extraction is carried out by convolution and maximum pooling, and then the ridge regression algorithm is used for emotion recognition. When the network needs to expand, the network is dynamically updated by incremental learning algorithm. We verified the experimental performance through k-fold cross validation. According to the recognition results, the accuracy on JAFFE database of our proposal is greater than that of the state of the art, such as the Local Binary Patterns with Softmax and Deep Attentive Multi-path convolutional neural network.

Keywords: Convolution Neural Networks, Broad Learning, Emotion Recognition, Human-Robot Interaction.

1. INTRODUCTION

With the rapid development of computational intelligence technology, more and more intelligent robots appear in human life, and people have higher and higher expectations for robots to have emotional abilities [Coelho (2017)]. However, the emotional capacity of current machines is insufficient to meet human needs [Luo (2019)]. Research shows that 55% of all human emotions are expressed through facial emotion [Tsalamlal (2017)]. Therefore, the realization of facial emotion recognition will help machines recognize human emotions [(Chen, 2017, 2020)] and even understand human emotional intentions.

Emotion recognition belongs to the research of pattern recognition and machine learning [Yin (2017)], so many recognition algorithms used in machine learning and pattern recognition are widely used in emotion recognition. However, existing machine learning methods always rely heavily on the representation of data [Zhang (2019)]. Representational learning provides a way to solve this problem. Although representations generally perform better than manual design, the learning representation is still unable to extract more abstract features. Deep learning [Hochreiter (1997)] realizes local correspondence through convolution operation and obtains the features of the whole picture through weight sharing. The problem of feature representation is solved [L. F. Chen (2018)]. At present, the most commonly used deep network includes deep Boltzmann machines (DBM) [Kim (2018)], deep belief networks (DBN) [Wang (2018)], recurrent neural networks (RNN) [C. L. P. Chen (2017)] and convolutional neural networks (CNN) [Li (2019)]. Although the deep network has been widely used and the application effect is obvious, the training and updating process of the network takes too long [L. F. Chen (2019)] because the depth structure usually involves a large number of parameters and the parameter optimization needs to adopt the gradient descent method to update step by step. The random vector function-link neural network (RVFLNN) can cover the disadvantages of long training time and satisfies the ability of the proximation of function. RVFLNN is proved to be a general approximation

⋆ This work was supported by the National Natural Science Foundation of China under Grants 61973286, 61603356, and 61733016, the 111 project under Grant B17040, and the Fundamental Research Funds for the Central Universities, China University of Geosciences (No. 201839).
† Corresponding author: Xuzhi Lai (e-mail: laixz@cug.edu.cn).

method for continuous function on a compact set in a wide range of control and modeling applications. Based on RVFLNN's ideas, Broad Learning (BL) [Y.-H. Pao (1992)] is established, which provides an alternative to machine learning by increasing the width of the network structure.When the network needs to add new feature nodes or enhance nodes, BL can quickly update the network weight through a matrix operation.

We used CNN to extract deeper and more abstract features in the feature extraction stage and reduced the feature dimension through maximum pooling. Then emotion recognition is realized through broad learning. Once new nodes need to be added, the network can be reconstructed efficiently through incremental learning.

CNN-based broad learning with efficient incremental reconstruction model (CNNBL) is proposed in this paper. CNN is used for feature extraction and selection. BL completes the emotion recognition and eventually produces facial emotion recognition results. To be specific, firstly, the convolution operation is used to realize the local feature response to obtain facial emotional information and generate the feature matrix. Considering the weak characterization ability of features, we adopt maximum pooling method to aggregate features and reduce feature dimensionality. Then, the ridge regression algorithm in the BL was used to complete the emotion classification. In addition, the incremental learning algorithm in BL is used to quickly update the network weight and achieve the highest recognition rate.

The main innovation of our proposal lies in the combination of deep learning and broad learning to facial emotion recognition. And in this paper, the main contribution is the use of CNN for extracting the deep and abstract features from the facial emotional images. At the same time, BL is adopted to realize fast updating of network weight, avoiding the problems of a complex network structure and slow updating speed. What's more, in order to reduce the time consumption of retraining, we adopted incremental learning, which can be rapidly reconstructed in large-scale expansion. The result of emotion recognition on JAFFE database proves the effectiveness of our proposal.

2. EMOTION RECOGNITION USING CNN-BASED BROAD LEARNING

Due to the complexity of facial features, facial emotion recognition faces certain difficulties and challenges. Different from the traditional machine learning method, CNN constructs a multi-layer deep neural network to abstract feature information layer by layer, so as to obtain more generalized and representational features. Therefore, we adopted CNN for feature extraction. Then BL is used for emotional recognition. The seven basic emotions are neutrality, happiness, anger, fear, surprise, sadness and disgust. Figure 1 is based on the structure of CNNBL with efficient incremental reconstruction model for facial emotion recognition.

2.1 Facial feature extraction using CNN

Most of the existing emotion recognition methods need to preprocess the input image, while CNN can directly take the image as the input of the network, which avoids the complicated data reconstruction process of traditional recognition methods. At the same time, the convolutional network structure is highly invariable and is not easily affected by other forms of deformation such as inclination, translation and scaling. We select CNN as the method of deep feature extraction in facial emotional images.

feature extraction using convolutional layer. Convolutional neural network is a deep neural network, which is composed of input layer, output layer and hidden layer. The hidden layer is mainly composed of pooling layer, convolutional layer and full connection layer. Among them, the convolutional layer mainly realizes the local feature response through convolution operation, and then uses the same convolution kernel to scan the whole image, extract the features of the whole image, and realize weight sharing. Generally, each convolution layer corresponds to multiple different convolution kernels, and the image features extracted by each convolution kernel are called feature graphs. The specific calculation is as follows:

$$y_j^l = \theta\left(\sum_{i=1}^{N_j^{l-1}} \omega_j \bigotimes x_i^{l-1} + b_j^l\right), j = 1, 2, ..., M. \quad (1)$$

where, y_j^l represents the j-th characteristic map; ω_j represents the corresponding convolution kernel; $\bigotimes$ for convolution operation; x_i^{l-1} represents the i-th feature map of the upper layer as the current input; b_j^l represents the bias; N_j^{l-1} represents the number of features of each feature map; M represents the number of characteristic graphs of each convolution layer; $\theta()$ is the activation function, the commonly used include tanh, sigmoid, ReLU, etc.

feature aggregation using pooling layer. In the convolutional neural network, the output of the convolutional layer is the feature of the image. The main function of pooling layer is to aggregate statistics of features, which aims to improve feature characterization ability and reduce the feature dimension. The most commonly used pooling algorithms in convolution neural networks include average pooling, random pooling and maximum pooling. The only difference between different pooling algorithms is that the convolution kernel selected is different. The operation reads as follows:

$$S = \sum_{i=1,j=1}^{c} p_{ij} x_{ij}. \quad (2)$$

where, c is the size and stride of the pooling domain, p_{ij} is the parameter of the operates, if it is average pooling algorithm, it takes $p_{ij} = \frac{1}{c^2}$; if it is maximum pooling, it takes 1 as the parameter of maximum eigenvalue, and the rest are all 0.

The role of the maximum pooling algorithm is to maximize the eigenvalues in the pooling domain, which can enhance the ability to depict emotional characteristics. The expression of maximum pool is as follows:

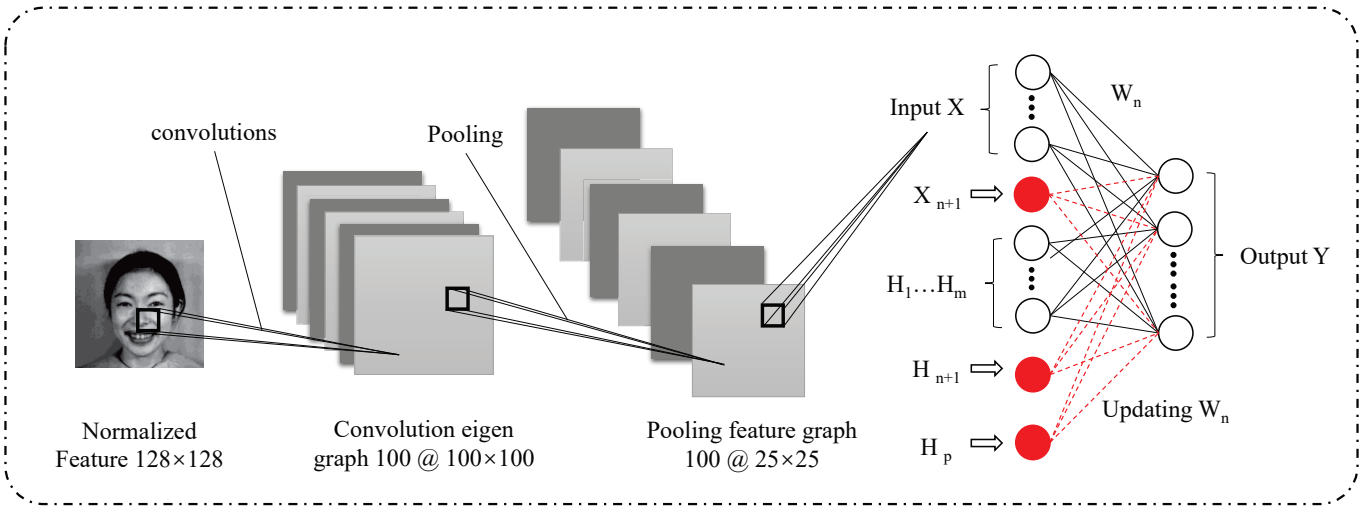

Fig. 1. The structure of CNNBL with efficient incremental reconstruction model for facial emotion recognition.

$$S_{i,j} = max_{i=1,j=1}^{c}(x_{ij}) + b \tag{3}$$

where, c is the size and stride of the pooling domain, and the feature graph after the pooling operation is matrix $S_{i,j}$.

2.2 Facial feature recognition using board learning

Broad learning can improve the speed of network training and provides an alternative way of deep learning network. The idea of BL design is as follows: First, the input data are mapped to features by mapping are used as the feature nodes of the network. Second, the features in the map are enhanced to enhancement nodes,which can generate the weights randomly. Finally, all of the mapped features nodes and enhancement nodes are Connected directly to the output. Through pseudo-inverse expression, the corresponding network weight can be obtained. BL can extend feature nodes and enhancement nodes in width. At the same time, if the network structure needs to be extended, the fast incremental learning algorithm can avoid the need for a complete network retraining.

The core of BL is to find the pseudo-inverse of feature node and enhancement node to target value. The pseudo-inverse matrix includes the weights of the neural network. The steps of forming a mapping of the input data to the feature node are as follows:

Step 1 : Normalize $H^{'}_{1(s\times f)}$ with z-score to ensure that the input data has been normalized to between 0 and 1.

Step 2 : At the end of training set $H^{'}_{1(s\times f)}$, add a column of value 1 to the matrix to expand it to $H_{1_s\times(f+1)}$, so that the bias term can be added directly through matrix operation when generating feature nodes.

Step 3 : Generate a random weight matrix w_e, which is a $(f+1)\times N_1$ dimensional random weight matrix with gaussian distribution.

Step 4 : Update W_{e_i} according to w_e, i represents the iteration quantity, and the number of iterations is N_2.

Step 5 : Update random eigenvector A_1=$H_1 \times w_e$. A_1 is normalized and sparsely autoencoded.

Step 6 : The feature node T_1 of a window is then generated as $T_1 = normal(H_1 \times W)$, *normal* represents normalization.

where $H^{'}_{1(s\times f)}$ stands for training set, s refers to the number of samples, f refers to the number of features, N_1 is feature nodes' numbers in each window, and N_2 refers to the number of windows for feature nodes. The eigennode matrix of the entire network is y.

Another characteristic of broad learning network is that the corresponding enhancement nodes in the network can be used to supplement the random feature nodes. As with feature nodes, the feature node matrix is firstly normalized and augmented to obtain H_2. However, being different from feature nodes, the coefficient matrix of enhanced nodes is not a random matrix, but a normalized random matrix.

Then the enhanced node is activated, and the matrix of the activated characteristic node is denoted as T_2.

$$T_2 = tansig(\frac{H_2 \times w_h \times s}{max(H_2 \times w_h)}) \tag{4}$$

where s is the scaling scale of the enhancement node, *tansig* is activation function in neural networks, the coefficient matrix of the enhanced node is w_h.

The final input T_3 of the network can be expressed as

$$T_3 = \begin{bmatrix} y \\ T2 \end{bmatrix} \tag{5}$$

The weight matrix W can be obtained by the ridge regression algorithm as follows:

$$argmin : ||T_3W - Y||_v^{\sigma_1} + \lambda||W||_u^{\sigma_2} \tag{6}$$

where $\sigma_1 > 0, \sigma_2 > 0$, and u, v are parameters in a typical standard regularization. The optimal problem is setted

by making $\sigma_1 = \sigma_2 = v = u = 2$, and using regular l2 norm regularization. It is convex function with a greater generalization performance. Y is the matrix of output. λ is a constraint on the square weight of W. When $\lambda = 0$, the solution to the original problem can be obtained by the least squares problem, instead of the inverse problem. On the other hand, the solution of the network is tends to 0 because of the highly constrained when $\lambda \to \infty$. Then, we have

$$argmin : W = (\lambda I + T_3 T_3^T)^{-1} T_3^T Y \tag{7}$$

When new data is added, we use incremental learning to obtain the new weight matrix W_1. The core of incremental learning is that, the updated weights can be obtained with only a small amount of calculation, using the newly added data and the last calculation result. The specific procedure is that the new column a is added to the previous input matrix T_3. Make $T_4^+ \triangleq [T_3|a]$. Then the new pseudo-inverse matrix for T_4^+ is

$$\begin{bmatrix} T_3^+ - db^T \\ b^T \end{bmatrix} \tag{8}$$

where $d = T_3^+ a$, $c = a - T_3 d$, if $c \neq 0$, $b^T = (c)^+$, if $c = 0$:

$$b^T = (1 + d^T d)^{-1} d^T T_3^+ \tag{9}$$

Finally, the weights of network are

$$W_1 = \begin{bmatrix} W - db^T Y \\ b^T Y \end{bmatrix} \tag{10}$$

Through the above method, when the new weights of the network need to be updated, it becomes only necessary to calculate the corresponding pseudo-inverse of the newly added nodes. It is important to note that if T_3 is full rank, then $c = 0$. The pseudo-inverse T_4^+ and the weight W_1 will be updated quickly.

3. EXPERIMENTS WITH CNN-BASED BROAD LEARNING

We verify the validity of the proposal on the JAFFE database. Meanwhile, the comparative experiment of incremental learning verifies the effectiveness for increasing network nodes. Finally, the facial emotion recognition is completed, and the results are analyzed.

3.1 Experimental environment and data selection

Figure 2 shows the emotion social robot interaction system we built for facial emotion recognition.

It consists of emotion computing workstations, data transmission equipment, routers and two mobile robots. The facial emotion workstation is equipped with a CPU of Intel i5-4590, which has a 3.3GHZ frequency of CPU, 4.00GB RAM and 64-bit operating system type. We selected MATLAB R2015b as the experimental software and the corresponding simulation experiments are designed

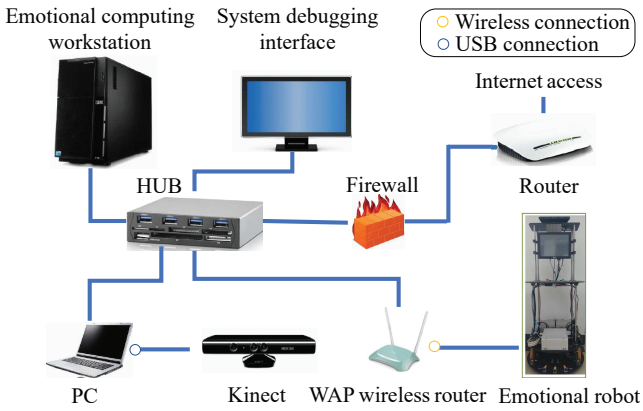

Fig. 2. The structure of facial emotion social robot interaction system.

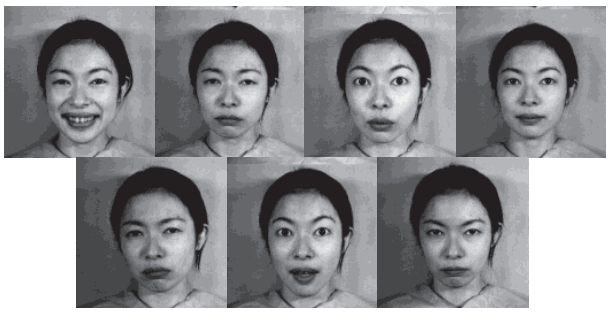

Fig. 3. Part of the sample image in the database.

for proving the validity of the algorithm. The JAFFE database is selected as the facial emotional database in our experiment.

The JAFFE database consists of seven basic expressions coming from 10 women, with a total of 213 gray-scale images. In JAFFE, each person has 2 to 4 images. 70% of the images are used as the training data, and the testing set consists of the remaining 30% of the data. The sample images of JAFFE database are shown as figure 3.

3.2 Simulations and analysis

In the feature extraction part, we adopted the convolution kernel of 29×29 with a stride of 1. The maximum pooling of 4×4 is adopted, and the stride is 4. In order to investigate the influence of incremental learning algorithm in broad learning on facial emotion recognition accuracy and experimental time, comparative experiments were designed. The facial emotion recognition results of the following experiments are obtained by the ten-fold cross validation. The training time, testing time and accuracy of the experiments were compared, and the comparison results are shown in table 1. FN means the feature nodes in CNNBL, and EN means the enhancement nodes corresponding to the feature nodes.

The confusion matrix of facial emotion recognition results of CNNBL is shown in figure 4. The recognition result of adding 15 feature nodes (FN) is shown in figure 5. When 15 enhancement nodes (EN) are added, the confusion matrix for facial emotion recognition result is shown as figure 6. When 15 enhancement nodes corresponding to the feature

Table 1. Facial emotion recognition results with node changes.

Node changes	Training time(s)	Testing time(s)	Accuracy(%)
CNNBL	5.0647	0.63290	87.14
CNNBL+15FN	5.9733	0.69384	88.57
CNNBL+15EN	6.8092	0.70818	92.85
CNNBL+15(FN+EN)	6.9890	0.74797	90.00

nodes and 15 feature nodes are added at the same time, figure 7 shows the confusion matrix of recognition result.

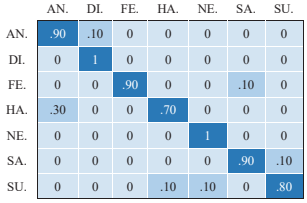

Fig. 4. Confusion matrix by CNNBL.

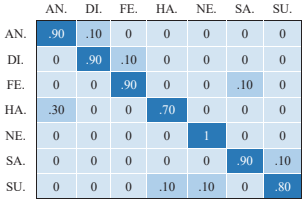

Fig. 5. Confusion matrix by CNNBL+15FN.

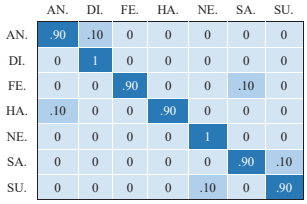

Fig. 6. Confusion matrix by CNNBL+15EN.

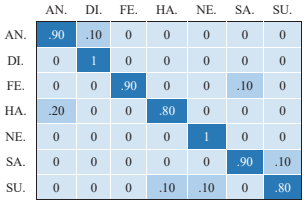

Fig. 7. Confusion matrix by CNNBL+15(FN+EN).

The facial emotion recognition results show that when keeping the number of input nodes and feature nodes unchanged and the enhanced nodes are added, the facial emotional recognition method obtains the highest recognition accuracy 92.85%. With the increase of the number of nodes, the training time and test time of CNNBL increased. In addition, we can conclude from the results of the last two experiments that the increase of enhanced nodes in the network has greater influence on the experimental results than the increase of feature nodes. Considering the influence of directly increasing input images instead of feature nodes, we added 15 images when the enhanced nodes in the experiment remained unchanged, and the recognition accuracy was 88.23%. The test time increased by adding pictures was only 0.4809s. We calculated the kappa coefficients of the above confounding matrices, and the kappa coefficients of the four experiments were 0.867, 0.850, 0.917, and 0.883, respectively, all in the range of 0.81 1, showing that the classification results were "almost perfect".

We compared our proposal with three other basic methods include local binary patterns with Softmax (LBP-Softmax), convolutional neural network and specific image processing steps (CNN-SIPS), and deep attentive multi-path convolutional neural network (DAM-CNN). The result of comparison is listed as Table 2. Compared with the three methods of facial emotional feature extraction and recognition mentioned above, the proposal achieves a higher recognition accuracy.

Table 2. Facial emotion recognition: comparative results.

Method changes	Accuracy(%)
LBP-Softmax [H. Ali (2018)]	91.27
CNN-SIPS [A. T. Lopes (2017)]	91.80
DAM-CNN [S. Y. Xie (2019)]	91.67
CNNBL	92.85

Moreover, our proposal takes into account the distortion of the input features, and can extract the high-level abstract features from the real images by using CNN. Incremental learning can quickly update the weights of the network as the number of inputs increases. At the same time, the depth and width of facial emotion recognition are combined. We will further investigate the effectiveness of our proposal on other facial emotion data sets.

4. CONCLUSIONS

The CNNBL is proposed for facial emotion recognition. On the premise of ensuring translation invariance, the extraction of facial emotional features is realized by convolution operation and maximum pooling. Broad learning algorithm is used for feature recognition. In addition, the network weight is updated rapidly by adopting the incremental learning method of generalized learning. The experimental results show that the proposal achieves a higher recognition accuracy than the above mentioned facial emotional recognition methods, such as LBP-Softmax, DAM-CNN, etc. Moreover, compared with existing deep learning algorithms, such as RNN, it has obvious advantages in terms of recognition time.

With the increasing demand for emotion recognition and intention understanding in human-computer interaction, it will be an interesting research topic for robots to recognize, understand and adapt to human emotions through behavior. In the future research, we'll further explore the methods of face emotion recognition based on width, depth and time scale, and complete the application experiment of our proposal to achieve more smooth human-computer interaction.

REFERENCES

J. P. Coelho, T. M. Pinho, and J. Boaventura-Cunha (2017). A new brain emotional learning simulink toolbox for control systems design. *IFAC-PapersOnLine*, 50 (1): 16009–16014.

Z. J. Luo, J. H. Chen, T. Takiguchi, and Y. Ariki (2019). Emotional Voice Conversion Using Dual Supervised Adversarial Networks With Continuous Wavelet Transform F0 Features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27 (10): 1535–1548.

M. Y. Tsalamlal, M. A. Amorim, J. C. Martin, and M. Ammi (2018). Combining Facial Expression and Touch for Perceiving Emotional Valence. *IEEE Transactions on Affective Computing*, 9 (4): 437–449.

L. F. Chen, M. Wu, M. T. Zhou, Z. T. Liu, J. H. She, and K. Hirota (2017). Dynamic emotion understanding in human-robot interaction based on two-layer fuzzy SVR-TS model. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, DOI: 10.1109/TSMC.2017.2756447.

L. F. Chen, W. J. Su, Y. Feng, M. Wu, J. H. She, and K. Hirota (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, 509: 150–163.

Z. Yin, Y. Wang, and L. Liu (2017). Physiological feature based emotion recognition via an ensemble deep autoencoder with parsimonious structure. *IFAC-PapersOnLine*, 50 (1): 6940–6945.

T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li (2019). Spatial-Temporal Recurrent Neural Network for Emotion Recognition. *IEEE Transactions on Cybernetics*, 49 (3): 839–847.

L. W. Kim (2018). DeepX: Deep Learning Accelerator for Restricted Boltzmann Machine Artificial Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (5): 1441–1453.

C. Y. Wang, J. C. Wang, A. Santoso, C. C. Chiang, and C. H. Wu (2018). Sound Event Recognition Using Auditory-Receptive-Field Binary Pattern and Hierarchical-Diving Deep Belief Network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26 (8): 1336–1351.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11): 2278–2324.

S. Hochreiter and J. Schmidhuber (1997). Long short-term memory. *Neural Computation*, 9 (8): 1735–1780.

L. F. Chen, M. T. Zhou, M. Wu, J. H. She, Z. T. Liu, F. Y. Dong, and K. Hirota (2018). Three-layer weighted fuzzy SVR for emotional intention understanding in humanrobot interaction. *IEEE Transactions on Fuzzy Systems*, 26 (5): 2524–2538.

C. L. P. Chen and Z. Liu (2017). An effective and efficient incremental learning system without the need for deep architecture. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (1): 10–24.

Y. Li, J. B. Zeng, S. G. Shan, and X. Chen (2019). Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism. *IEEE Transactions on Image Processing*, 28 (5): 2439–2450.

L. F. Chen, Y. Feng, M. A. Maram, Y. W. Wang, M. Wu, K. Hirota, and W. Pedrycz (2019). Multi-SVM based Demp-ster-Shafer Theory for Gesture Intention Understanding Using Sparse Coding Feature. *Applied Soft Computing*, DOI: 10.1016/j.asoc.2019.105787.

Y. -H. Pao and Y. Takefuji (1992). Functional-link net computing: Theory system architecture and functionalities. *Computer*, 25 (5): 76–79.

B. Igelnik and Y. -H. Pao (1995). Stochastic choice of basis functions in adaptive function approximation and the functional-link net. *IEEE Transactions on Neural Networks*, 6 (6): 1320–1329.

H. Ali, M. Hariharan, S. Yaacob, and A. H. Adom (2018). Facial Expressions Recognition Based on Cognition and Mapped Binary Patterns. *IEEE Access*, 6: 18795–18803.

A. T. Lopes, E. D. Aguiar, S. F. D. Souza, and T. O. Santos (2017). Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognition*, 61: 610–628.

S. Y. Xie, H. F. Hu, and Y. B. Wu (2019). Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognition*, 92: 177–191.