Facial Emotion Recognition Using Convolution Neural Networks-Based Deep Learning Model

Chinonso Paschal Udeh[1,2,3], Luefeng Chen[1,2,3, *], and Min Wu [1,2,3]

[1]School of Automation, China University of Geosciences, Wuhan 430074, China
[2]Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex System, Wuhan 430074, China
[3]Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, China
E-mail: chenluefeng@cug.edu.cn

Abstract. With the development of emotion recognition, learning, and analysis, robotics plays a significant role in human perception, attention, memory, decision-making, and social communication, leading to emotion recognition and human-robot interaction (HRI). This research analyzes the interaction between humans and robots using facial expressions and head pose to achieve robustness in understanding emotions by optimizing the traditional deep neural networks to comprehend the coexistence of multi facial information in HRI using convolution neural networks. A hybrid genetic algorithm with stochastic gradient descent is adopted, which has the capacity of inherent, implicit parallelism and better global optimization of the genetic algorithm to find the better weights of the network. The experiment shows the proposal's effectiveness in providing complete emotion recognition through single-modal cooperation of HRI that can interact with humans and machines.

Keywords: convolution neural network, deep learning, emotion recognition, facial expression, head pose.

1. INTRODUCTION

Numerous scholars have studied facial emotion recognition (FER) over the years [1]. The quest for sufficient understanding towards its stability and optimization for a smooth interaction between humans and robotics draws global attention as various nations seek to improve the already existing technology. Recently, FER has been the main focal point regarding emotion recognition, giving that facial emotion alone accounts for about 55% of regular communication compared to the other affective computing [2]. Even though human beings express thousands of different facial expressions, words and facial expressions are different in application. While words can have precise meanings, facial expressions have numerous meanings in general [3]. While reading the face, focusing on emotions is recommended, rather than figuring out the facial expression [4, 5]. For "Honestly, I did not take the last Apple from the fridge." Emotions mainly focus on the seven central affective states [6, 7] angry, happiness, neutral, sadness, surprise, disgust, and fear. That has been scientifically documented to be associated with specific facial configurations [8, 9]. These emotions are called basic emotions, and facial expressions of these emotions can be reliably recognized across all cultures worldwide. [10] warned that facial expression recognition should not be misinterpreted for human emotional detection. However, since the emotional state activates the facial muscles, facial expression analysis could detect human emotion.

Facial emotion recognition has gained popularity in affective computing. The human face conveys more information in nonverbal communication channels than speech, gesture, and other counterparts [11]. The seven emotions listed above are not the only universally experienced emotions but are just the ones with widely recognized expressions. If someone wants to be more accurate at reading facial expressions, focus on recognizing the basic emotions. By examining the entire face, facial expressions can involve multiple areas of the face. For example, the fundamental movements commonly associated with surprise are: The whole eyebrows are pulled up, the upper eyelids are pulled up, the jaw drops or is pulled down. Another common mistake is to assume an expression of emotion without actually seeing all of the movements. If you see the eyebrow-raising movement described above, it could be a partial expression of surprise. However, the eyebrows are also raised in fear (and combined with a third eyebrow movement). Beyond emotional words, eyebrow-raising is also commonly used to accent parts of speech. Stability and increasing demands on FER accuracy indeed depend on focusing on the changes across the entire face. Facial emotion recognition involves looking for movements in the whole face, for higher accuracy in recognizing facial expressions of emotion involved [12], the HRI ability to catch what we humans do not see as well as what we do [13].

This paper presents a hybrid genetic algorithm with stochastic gradient descent (HGASGD) by minimizing the classification error and training the CNN classifiers to find a local minimum during fitness evaluations of genetic algorithm (GA) chromosomes. The experiment showed that by combining the evidence of classifiers generated with the help of GA's leading to improved performance, the GA's optimized the weight of CNN. A hybrid GA with stochastic gradient descent (SGD) is developed. The weights

optimized by SGD will be taken as a chromosome of GA's initial population to accelerate the algorithm's convergence.

The rest of this paper is as follows. In section 2, facial expression and head pose means of extraction are reviewed and discussed. Section 3 explained the CNNDL model for emotion understanding. In section 4, the experimental results and analysis are presented along with a discussion.

2. RELATED WORKS

Facial Action Coding System (FACS) is a framework created by Ekman et al.[14] to quantify the activity of facial muscle movements. FACS framework defines Action Unit (AU) as a numeric code that describes facial muscle activity [15], e.g., cheek raising, brow lowering, wrinkling of the nose, raising of upper lips, etc. AU is anatomically related to facial muscles' action that explains almost all possible facial muscle movements. There are 46 AUs on a face, and those are used as the essential component to compose a facial expression; the AU arrangement is usually different for each emotion [16]. Therefore, decomposing the facial expression into the AU set is an essential step in analyzing facial expression. AU recognition continues to be an active research area due to the complexity and challenges it posed. Two approaches in AU recognition; occurrence-based and intensity-based estimation [17]. FACS defines five AU intensities, from level A to level E, where the higher the intensity value, the firmer its occurrence [17]. The advantage of AU intensity estimation is that the information that is generated by the observed subject is rich and deep. AU combination and translates it into respected basic emotion classes.

AU is effectively used as an attempt to quantify emotions on a face area. The use of AU makes it easier to translate a facial expression into a relevant feeling. Another approach besides AU is simply by using Facial Points (FP) analysis. FP is landmarked on a face used as control points and located on essential face elements such as eyebrows, eyes, nose, and mouth. However, since facial expressions vary, facial feature extraction methods must be considered to get the best recognition results.

3. CONVOLUTION NEURAL NETWORKS-BASED DEEP LEARNING

In neural networks theory, obtaining accurate emotional information is an essential factor in understanding emotional intention. It is also the key to promote harmonious communication between man and machine. In the process of dynamic interaction, emotion is mainly expressed by facial expressions and head poses. Therefore, to obtain accurate emotional information, an emotion recognition algorithm based on single modal surface communication information should provide accurate dynamic information for subsequent emotional understanding. Different modal report with other properties and different structure design of convolution neural network to extract the facial expressions and high-level dynamic characteristics of voice information data, and then the modal based on canonical correlation analysis fusion dynamic characteristics to extract more information from multiple modals said discriminant features, to obtain more accurate emotional details.

3.1 Emotion Expression Based on CNNDL

For low-level emotional features extracted to CNNDL model is set up so that the characteristics of different modal of low-level emotional design different structure of the neural network to extract of high-level abstract emotional aspects, and the modal based on canonical correlation analysis algorithm fusion high-level emotional features to extract more information from multiple modals said identifying characteristics. Then, the Softmax regression model was built on the top of the network to recognize the fused emotional features and realize the emotional recognition based on the modal surface information.

Its application methods are given the below:

(i) For n different types of inputs, we construct n CNNs to extract n different kinds of features which the entire connection layer is followed by the SR layer in each CNN to form a complete network.
(ii) SGD is obtained to optimize the CNNs.
(iii) The features generated by each CCNs are fused in the feature fusion layer, where the corresponding features generated by each CNNs are added.
(iv) The fusion features are input to the Softmax layer for classification, and then the CNNDL is formed.
(v) Use HGASGD to optimize the CNNDL.
(vi) The CNNs are composed of convolutional layers and sub-sampling layers iteratively to extract high-level semantic features, a dropout layer to alleviate the overfitting, and a complete connection layer.

The convolution layers use the convolution kernel to implement the local connections between neurons in the neighboring layer to excavate local association information in the PCA feature maps, excavating high-level emotion features. The convolution process is implemented by

$$x_j^{(l+1)} = f(\sum_{i=M_j} x_i^l \times W_{ij}^{(l)} + b_j) \quad (1)$$

where $x_j^{(l+1)}$ is the i-th feature map data of l+1-layer, $\sum_{i=M_j} x_i^l \times W_{ij}^{(l)}$ is the convolution operation, summation for all the features map data of l-layer and the 1-th layer of the j-th convolution kernel W, and b_k is the deviation corresponding to the k-th convolution kernel. $f(\cdot)$ is a ReLu type activation function.

The convolution feature map generated by the convolution layer is then input to the sub-sampling layer for average pooling, weakening the network's sensitivity to the change of the emotion feature position in the feature map. The average pooling implementation formula is

$$x_j^{(l+1)} = f(x_i^l \times \frac{1}{p^2}) \quad (2)$$

where x_i^l is the i-th convolution feature map in the l-layer, and p is the average pooling dimension.

The dropout layer can effectively solve the overfitting problem of the DNN and improve the efficiency of feature

learning. The training phase randomly lets all input layer nodes work at each iteration, but their weights will be preserved; the calculation process is shown in (3). In the test phase, all network neurons are activated to make the network into a complete structure, equivalent to collecting multiple neural networks, whose calculation process is shown in (4).

$$D_{Train}(x) = \text{RandomZero}(p) \times x \quad (3)$$

$$D_{Test}(x) = (1-p) \times x \quad (4)$$

SR is the output layer of the network to classify the learned emotion feature. For the training set $\{(x^{(1)}, y^{(1)}), ...(x^{(m)}, y^{(m)})\}$, there is $y^{(i)} \in \{1, 2, ..., k\}$ .SR uses a cost function to evaluate its classification effect in a training phase; the definition of the Softmax cost function is

$$J(\theta) = -\frac{1}{m}\left[\sum_{i-1}^{m}\sum_{j=1}^{k} 1\{y^i = j\} \log \frac{e^{\theta_j^T x^i}}{\sum_{l=1}^{k} e^{\theta_j^T x^i}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2 \quad (5)$$

where $1\{y^{(i)} = j\}$ is an exponential function, the value of the rule is $1\{a \text{ true statement}\} = 1$ or $1\{a \text{ false statement}\} = 0$. The derivation of (5) is

$$\nabla_{\theta_j} J(\theta) = \frac{-\sum_{i=1}^{m}\left[x^{(i)}(1\{y^{(i)} = j\} - p(y^{(i)} = j \mid x^{(i)}; \theta))\right]}{m} + \lambda\theta_j \quad (6)$$

When training the SR, the (6) is substituted into the HGASGD to be minimized. The concrete implementation flow will be described in detail in section 3.2.

3.2 A Hybrid Genetic Algorithm with Stochastic Gradient Descent for Optimizing.

The HGASGD is proposed to optimize the CNNDL for global search, whose application methods are as follows

(i) Optimize each SCNN in CNNDL by SGD and preserve the optimized weights of CNNDL.
(ii) Initialize the population randomly, one of which is the weights obtained in (i).
(iii) Evaluate the fitness of each chromosome in the population by the fitness function.
(iv) Select chromosomes according to the fitness values by using stochastic universal sampling.
(v) Perform genetic operations of crossover and mutation to produce a new population, add the chromosome which has the best fitness values to the new people.
(vi) Repeat following (iii) to (v) iteration until the maximum number of generations is reached, and the best chromosome is obtained.

The weights optimized by SGD will be taken as a chromosome of GA's initial population in the proposal. The remaining chromosomes are obtained by random initialization. SGD update is given by

$$v_t = \gamma v_{t-1} + \alpha\nabla_\theta J(\theta; x^{(i)}, y^{(i)}) \quad (7)$$

$$\theta = \theta - v_t \quad (8)$$

where $\alpha\nabla_\theta J(\theta; x^{(i)}, y^{(i)})$ is the gradient of the θ calculated from the training set $(x^{(i)}, y^{(i)})$, α is a learning rate that prevents a large offset in the cost function. v is the current velocity vector, which is the same as the dimension of θ . $\gamma \in (0,1]$ determines how many gradient iterations will be involved in the update.

The chromosome representations of GA are real-coded. At each genetic iteration, the fitness evaluation is needed first, which is defined as (9) where $J(\theta)$ is defined as (6). Then, perform genetic operations of selection, crossover, and mutation to produce a new population. However, the chromosomes with the best fitness will be directly passed on to the new people.

$$\text{Fitness} = J(\theta) \quad (9)$$

Significantly to improve the optimization efficiency of GA, the stochastic mini-batch is applied in the GA to reduce the use of computing resources. And the data of the training set will be randomly shuffled before each epoch. This stochastic mini-batch method is also used in SGD.

4. EXPERIMENTATION

This analysis was conducted by making use of an affective computing workstation configured to make use of an Intel Core i5-4590 CPU processor possessing a 3.3GHz system clock and 6.0GB RAM. The analytical software adopted for this research is MATLAB R2018b. This analysis benchmark database is designed to confirm how practical nature of the proposal databases.

4.1 Databases

A bimodal face and head pose (FABO) database [19] is used in the experiment. In the FABO database, each video has four temporal phases: neutral, onset, apex, and offset. The apex phase presents the most significant features for emotion determination. Based on this fact, only the frames from the apex phase of each emotion are selected. This database contains a total of 10 emotional states (i.e., anger, anxiety, boredom, disgust, fear, happiness, puzzlement, sadness, surprise, and uncertainty) performed by 23 subjects of different nationalities, the partial facial expression and head pose frames from the apex phase of each emotion.

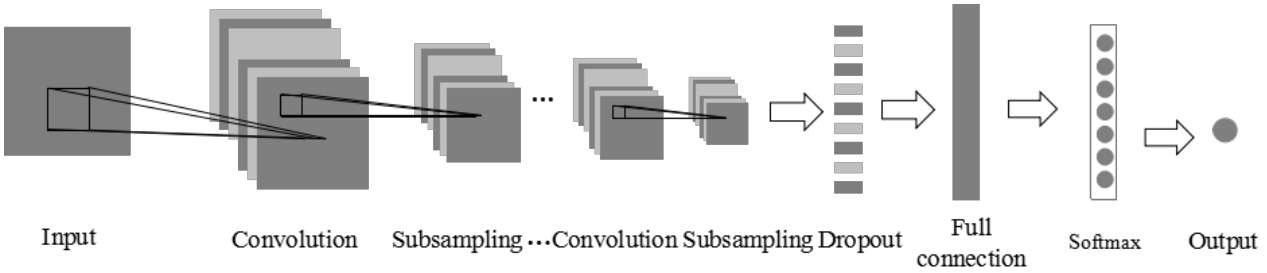

Fig. 1 Convolution neural networks-based deep learning model for emotion understanding

Moreover, the frames of neutral state are also selected, leading to 11 emotional states to be classified. Therefore, a total of 649 facial expression frames and 649 head pose frames are selected. Select 80% of facial expression samples and head pose samples for the training set, another 20% for the test set.

Three experiments are executed and evaluated. The first one uses information of the facial expression to determine emotional states. The second one extracts data from the head pose, and the third one uses both modal information. And the CNNDL in the experiments use HGASGD to optimize, the CNNs in the experiments use SGD to optimize. The experimental results are shown in Table 1, which also gives the experimental results of other state-of-the-art models. It can be seen that the CNNDL achieves better accuracy than that of CNNs. The average accuracy of CNNDL is 82.14% on facial expression, 82.14% on the head pose, and 91.43% on modal data of facial expression and head pose. Fig. 2 shows the confusion matrices of CNNDL on modal data of facial expression and head pose, where the accuracy of disgust, happiness, and uncertainty are 100%.

Table. 2 Comparison of CNNDL with other states of the art models

Recognition index	Facial Expression	Head Pose	Both
CNNDL	82.14	80.71	91.43
CNNs	79.29	78.57	88.57

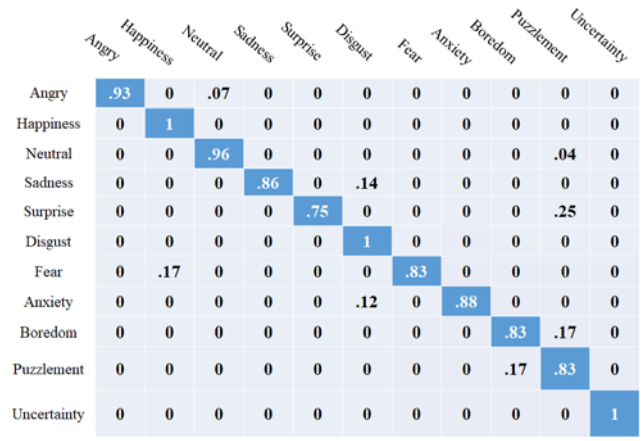

Fig. 2 Confusion matrices of CNNDL

4.2 Layer Structure

There are two kinds of modal data, namely, facial expression data and head pose data. The Viola-Jones face detection algorithm [20] is used to extract the face area for facial expression frames. Then the extracted face expression and posed frames are normalized to a uniform size 48×48 and 32×48, respectively. As for the CNNs of facial expression, there are two convolution layers and two average pooling layers. The size of the convolution kernel in the two convolution layers are 25×25 and 7×7, respectively, and the number of them is 40 and 50, respectively, average pooling dimensions are both 2. For the CNN of head pose, there are also two convolution layers and two intermediate pooling layers. The size of the convolution kernel in two convolution layers are 15 × 25 and 6 × 7, respectively, and the number of them is also 40 and 50, respectively, average pooling dimensions are both 2. The *p* in the dropout layer is 0.5 in both SCNNs.

4.3 Results and Discussion

The proposed modal data was executed through configuring the human-robotic system, then transmits the data to the system first acquires modal data through Kinect configured on the HRI system, then sends the data to the affective computing workstation. Then, the workstation will input the data into the trained CNNDL to identify the emotional states. The datasets of both facial expression and head pose after been fussed out results in much better output, after multiple iterations on a single GPU for about 80 seconds. The convergence line of fitness is well displayed below respective towards both accuracy and the loss in Fig.3 and 4.

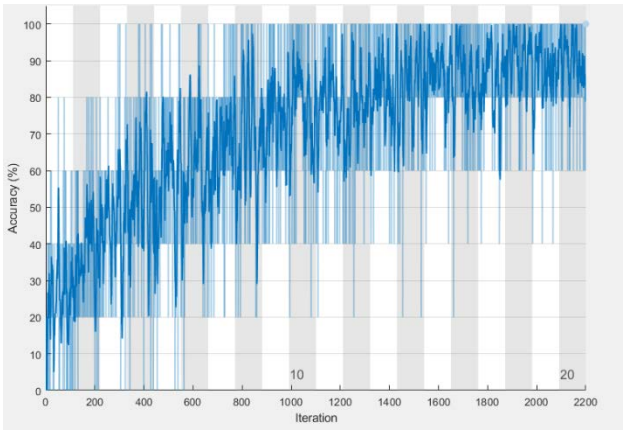

Fig. 3 Accuracy of different iterations

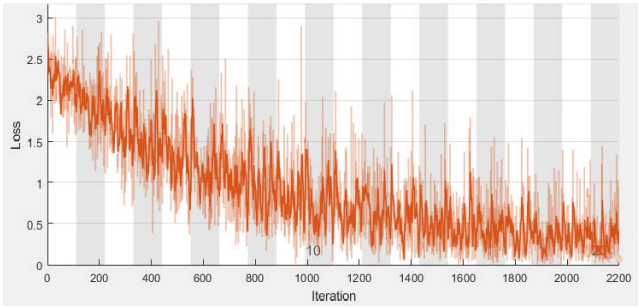

Fig. 4 Loss convergence line of fitness

The emotion understanding information will be feedback to the HRI system to achieve natural and harmonious interaction. This experiment is carried out in a working environment and analyzed volunteer's intention to voluntarily pose towards the camera for passport head shoot and voluntary and involuntary photo shooting when they are in random emotions in the working environment. According to our previous research [18], students' behavior intentions differ in different feelings. So, the human-robotic system will choose expression for the volunteer-based on the volunteer's emotional states, gender, the preference of different sexes for presentation in other emotions. Two student volunteers from our team participated in training and testing our models inside our school laboratory. The seven primary emotional states' frames (i.e., angry, happiness, neutral, sadness, surprise, disgust, and fear) of two student volunteers are collected by robot system as

training sets, including modal data of facial expression and head pose. Fig. 5 gives the confusion matrix of the preliminary application experiment of the two volunteers. The average recognition accuracy of the primary application experiment is 82% showing the application prospect of the proposal.

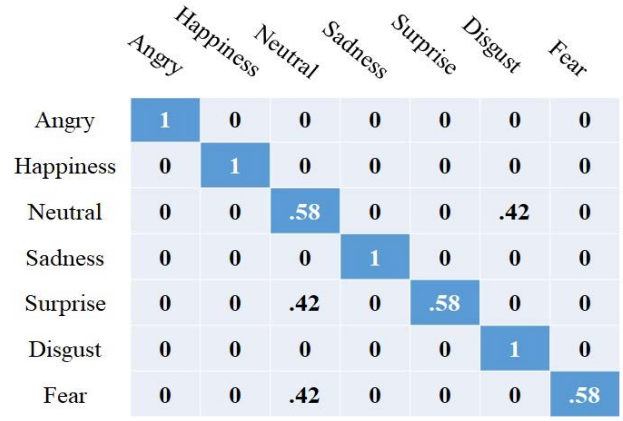

Fig. 5 Confusion matrix of application experiment

5. CONCLUSION

A convolution neural networks-based deep learning model combined with a dataset for emotion understanding is processed in this paper. The modal data of facial expression and head pose are used to recognize emotional states for emotional understanding. Adding the head pose, driven from the AU features, deep semantic information of emotional features fuses emotional facial expressions and head pose. Effectively increase the accuracy of the emotional state by developing HGASGD to improve the accurate and efficient search abilities for optimizing CNNs. The experiment showed that the proposed outperformed just the facial expression and the state-of-the-art methods.

In the future, we plan to extend the application experiment towards human-robot interaction by adding other modal features to increase emotional intention and offer new ideas and practical schemes for the natural harmonious to human-robot interaction.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China supports this work 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 under Grant 2021063.

REFERENCES:

[1] L Chen, M Wu, M Zhou, Z Liu, J She, K Hirota. Dynamic emotion understanding in human-robot interaction based on two-layer fuzzy SVR-TS model, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 50 (2): 490-501, 2020.

[2] J Call, M Carpenter. Three sources of information in social learning, Imitation in Animals and Artifacts, The MIT Press, 211-228, 2002.

[3] W M Tomasello. The cultural origins of human cognition, Harvard University Press, 2009.

[4] A L Thomaz, C Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners, *Artificial Intelligence*, 172 (6-7): 716-737,2008.

[5] N Ratliff. Learning to search: Structured prediction techniques for imitation learning, Carnegie Mellon University, 2009.

[6] R Toris, D Kent, S Chernova. The robot management system: a framework for conducting human-robot interaction studies through crowdsourcing, *Journal of Human-Robot Interaction*, 3 (2): 25-49, 2014.

[7] J Tao, T Tan. Affective computing: A review, *Proceedings of the International Conference on Affective computing and intelligent interaction*, 981-995, 2005.

[8] C Korsmeyer. Rosalind W Picard. Affective computing, *Minds & Machines*, 9 (3): 443-447,1999.

[9] R W Picard. Affective computing, The MIT Press, 2000.

[10] B Fasel, J Luettin. Automatic facial expression analysis: A survey, *Pattern Recognition*, 36 (1): 259-275, 2003.

[11] N Elfaramawy, P Barros, G I Parisi, S Wermter. Emotion recognition from body expressions with a neural network architecture, *Proceedings of the 5th International Conference on Human Agent Interaction*,143-149, 2017.

[12] M Soleymani, M Pantic, T Pun. Multimodal emotion recognition in response to videos, *IEEE transactions on Affective Computing*, 3 (2): 211-223, 2011.

[13] A T Lopes, E De Aguiar, A F De Souza, T Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, *Pattern Recognition*, 61: 610-628,2017.

[14] S J Vick, B M Waller, L A Parr, M C S Pasqualini, K A Bard. A cross-species comparison of facial morphology and movement in humans and chimpanzees using the facial action coding system (FACS), *Journal of Nonverbal Behavior*, 31 (1): 1-20, 2007.

[15] J Yan, Z Lei, L Wen, S Z Li. The fastest deformable part model for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2497-2504,2014.

[16] R Cowie, E Douglas-Cowie, J G Taylor, S Ioannou, S D Kollias. An intelligent system for facial emotion recognition, *Proceedings of the IEEE International Conference on Multimedia and Expo*, DOI: 10.1109/ICME.2005.1521570, 2005.

[17] H Li, Z Lin, X Shen, J Brandt, G Hua. A convolutional neural network cascade for face detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5325-5334, 2015.

[18] L Chen, M Wu, W Su, K Hirota. Multi-convolution neural networks-based deep learning model for emotion understanding, *Proceedings of the 37th Chinese Control Conference*, 9545-9549, 2018.

[19] A bimodal face and body gesture database, 2006, URL: http://mmv.eecs.qmul.ac.uk/fabo/.

[20] P Viola, M J Jones. Robust real-time face detection, *International Journal of Computer Vision*, 57 (2): 137-154, 2004.