Real-time multimodal emotion recognition system based on elderly accompanying robot

Shaosong Dou¹,², Zhiquan Feng¹,²*, Xiaohui Yang¹,², Jinglan Tian¹,²

¹College of Information Science and Engineering, University of Jinan, Jinan, Shandong, 250022, China
²Shandong Provincial Key Laboratory of Network Environment Intelligent Computing Technology, University of Jinan, Jinan, Shandong, 250022, China

*Corresponding author’s e-mail: ise_fengzq@ujn.edu.cn

Abstract. In recent years, people have been paying more attention to the impact of technological development on human emotion recognition. At the same time, China has become the country with the most elderly population in the world. However, due to the lack of real-time multimodal emotion recognition technology for the elderly accompany robots, this paper proposes a deep learning decision-level fusion of real-time emotion analysis model which is based on the background of the elderly care. The results of image and audio recognition are used for intersection and union operation to get emotional classification result, and the obtained emotional result is corresponding to the feedback behavior of the accompany robot. And after the experiment, the recognition algorithm proposed in this paper accuracy can reach about 90%, which is nearly 10% higher than the single mode, and the feedback of the robot has achieved the expected effect.

1. Introduction
Emotion recognition detection technology has become the focus of the current artificial intelligence field. First, facial expressions contain rich human emotion information, which plays an important role in human-computer interaction and pattern recognition. The speech recognition technology is gradually becoming a focus for detecting human emotions. However, relying on single-modal analysis can not achieve very accurate emotion recognition, so this paper proposes a fusion image recognition and audio recognition model for more accurate emotion recognition.

At the same time, the research results show that robots have the ability to reduce people's stubborn, can encourage the elderly to actively participate in rehabilitation therapy, and increase the sociality of the elderly[1]; the current elderly living alone are more and more, the mental health problems of the elderly become a problem, but their children can not accompany the elderly in time, accompany robot take real-time emotional recognition of the elderly, then it can take appropriate ways to improve the mood of the elderly.

2. Related work

2.1. Speech emotion recognition
Earlier, feature extraction in speech emotion recognition often used classical machine learning algorithms, such as hidden Markov models, support vector machines, or decision tree methods[2]. Due
to the neural network based on deep learning advance rapidly in recent years, the use of neural network models for speech emotion recognition has been greatly developed. [3] proposed recurrent neural network automatic speech emotion recognition can provide more accurate prediction results than the previous speech emotion recognition algorithm. It proposes a new collection strategy that uses BiLSTM and attention mechanism to enable the network to focus on emotionally prominent parts.

2.2. Expression recognition
At present, the application of neural network model to extract facial expression features and classification is the hot topic of current face recognition, [4] proposed an extended deep neural network for facial emotion recognition technology, it is a new deep fully connected model for facial emotion recognition; [5] proposes a deep convolutional neural network model, and experimental results on the FER-2013 data set show that this model has the same accuracy with the latest method, and in addition, this model uses fewer parameters than the latest model.

2.3. Multimodal fusion
For decision-level integration, the previous works focus on four typical fusion strategies: feature-level fusion, decision-level fusion, fractional-level fusion and model-level fusion. Feature-level fusion is the most common and direct way. All extracted features are directly concatenated into a single high-dimensional feature vector; fractional-level fusion is enhanced by combining individual classification scores. Model-level fusion aims to obtain a joint feature representation of audio and visual modes; the decision-level fusion aims to combine multiple single emotion recognition results through algebraic combination rules, and then these single mode recognition results are combined with certain algebraic rules[6].

3. Algorithm design

3.1. Multimodal fusion architecture
The multi-modal fusion model in this paper contains two separate input streams, namely the Xception[7] model based on convolutional neural network for processing vision, and the BiLSTM[8, 9] model based on RNN cyclic neural network for processing audio. Then, use the classification results obtained by the two networks to subject to intersection and complement operations.

![Figure 1. Multi-modal emotion recognition system flow chart.](image)

3.2. Algorithm description
The system of Figure 1 adopts a decision-level fusion scheme. The system is divided into an intent-aware layer, a multi-modal fusion layer and a quadratic behavior matching layer. The results obtained by the lightweight Xception network trained model and the BiLSTM network will be
intersected to get a result set. If the result set is not empty, the correct result is output, and If it is an empty set, perform the secondary behavior match.

Table 1 is the specific fusion algorithm design of the system. The expressions and speech recognition results belong to two sets with four elements, namely \( X = \{ 1: \text{angry}, 2: \text{sad}, 3: \text{neutral}, 4: \text{happy} \} \), \( Y = \{ 1: \text{angry}, 2: \text{sad}, 3: \text{neutral}, 4: \text{happy} \} \). They are set to the first two elements as negative emotions, and the last two are positive emotions; \( Q = \{ X \cap Y \} \) indicating that the intersection of \( X \) and \( Y \); in the second match, the negative emotion set is \( a = \{1, 2\} \), and the positive emotion set is \( b = \{3, 4\} \), and the \( p = \{ X \cup Y \} \) means the union of \( X \) and \( Y \). The matching algorithm is shown in Table 1. Than the obtained result is matched with the robot's intention recognition action set named \( Z \), and the action set \( Z \) has four actions, as shown in Table 2.

Table 1. Description of multi-modal fusion algorithm.

| Multi-modal fusion algorithm |
|-------------------------------|
| **Input:** Input the user's facial image information into the Xception network model in real time; input the user's three seconds of voice into the BiLSTM network model in real time every three seconds; | |
| **Output:** Voice prompt: You will activate the real-time emotion recognition program to ensure the safety of the surrounding environment of the robot; | |
| While (The robot ends the feedback action, or the recognition result is wrong.) | |
| { | |
| (1) Firstly, the expression recognition is performed to obtain the result \( X \), and if the user's facial expression recognition is obtained, execute (2); if the result is not obtained, execute (1); | |
| (2) Real-time obtain three seconds of user speech for mood emotion judgment to obtain result \( Y \); | |
| (3) The recognition results \( X \) and \( Y \) obtained by (1) and (2) are subjected to intersection calculation \( Q = X \cap Y \), and if \( Q \) not an empty set, then the correct result will sent to robot and it will perform motion feedback; if \( Q \) is an empty set, execute (4); | |
| (4) Set negative emotion result set \( a = \{1, 2\} \) and positive emotion result set \( b = \{3, 4\} \). Then set \( p = X \cup Y \) as the union of \( X \) and \( Y \) and match with the two emotion sets, if \( p \in a \) then judged as positive emotion, and execute (5); if \( p \in b \) then judged as positive emotion, and execute (5); if \( p \notin a \) and \( p \notin b \), it means that the two channels belong to positive and negative emotions respectively, which is inconsistent with the actual, so that is recognition error, then re-executed (1); | |
| (5) Comparing the probability values of the two channel results, probability \( X \) and probability \( Y \), to obtain the highest probability value as the correct result; | |
| (6) The robot matches the result with the action set \( Z \), and the robot performs corresponding actions; | |
| (7) End of the action, continue (1); | |
| } | |

End

Table 2. Robot's intention recognition action set \( Z \).

| Recognition Result | Robot Action |
|--------------------|--------------|
| Angry              | Robot voice prompt: You look a little angry, I will help you get a cup of water; then execute the water operation; |
| Sad                | The user's mood is sad, and the robot greets the user to communicate; |
| Neutral            | Robot voice prompt: You look a little sad, I will open the TV for you; then perform the TV operation; |
| Happy              | Robot voice prompt: You look like very happy, I will give you a dance; then execute the dance operation; |
4. Experiment

4.1. Operating environment
The CPU selected at the host during the experimental operation is Intel(R) Core(TM) i7-9750H @ 2.6 GHz. The Graphics card selected the Inter(R) UHD Graphics 630 NVIDIA GeForce GTX 1650.

4.2. Experimental process
If the final recognition result is angry, robot takes the water operation as (a); if the final recognition result is sad, robot takes the greet operation as (b); if the final recognition result is neutral, robot takes the TV operation as (c); if the final recognition result is happy, robot takes the dance operation as (d).

(a) water operation  (b) greet operation  (c) TV operation  (d) dance operation

Figure 2. Robot interaction example pictures.

4.3. Comparative experiment
In this paper, single-modal emotion recognition was selected as the control experiment of this experiment. Five elderly people were selected for this experiment, aged between 50 and 60 years old. The trials were divided into three groups as 50 experiments, 100 experiments and 200 experiments. Each of the elderly performed equal number of experiments. The recognition results were recorded as table 3; the obtained experimental accuracy comparison chart as shown in figure 3.

| Identification type | 50 correct times | 100 correct times | 200 correct times |
|---------------------|------------------|-------------------|-------------------|
| Expression          | 43               | 80                | 166               |
| Speech              | 37               | 74                | 157               |
| Multimodal          | 45               | 91                | 188               |

Table 3. Identify correct statistical table.

![Recognition Accuracy Chart](image)

Figure 3. Accuracy comparison chart.

5. Conclusion
This paper proposed a multi-modal emotion recognition algorithm based on deep learning for the elderly accompany robot. It can be seen from the experimental results that the accuracy of the multi-modal algorithm is better than single mode recognition. And the accuracy can reach about 90% which is nearly 10% higher than the single mode. In the background application of the accompany
robot, the expected results were obtained. The improvement of this paper is to improve the recognition rate of single mode, and improve the accuracy of the multimodal algorithm and robot actions.

Acknowledgments
This paper is supported by the National Key R& D Program of China (No. 2018YFB1004901), the National Natural Science Foundation of China (No. 61603151) and the Natural Science Foundation of Shandong Province (ZR2018PF012).

References
[1] Noguchi, H., Koyano, Y., Mori, H., Komiyama, C., Sanada, H., Mori, T. (2019) Exploration of communication robot use for older patients in an acute hospital based on case trials. Journal of Nursing Science and Engineering., 6(2): 70-82.
[2] Lin, Y. L., Wei, G. (2005) Speech emotion recognition based on HMM and SVM. In: 2005 international conference on machine learning and cybernetics. Guangzhou, China. pp. 4898-4901.
[3] Mirsamadi, S., Barsoum, E., Zhang, C. (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. pp. 2227-2231.
[4] Jain, D. K., Shamsolmoali, P., Sehdev, P. (2019) Extended deep neural network for facial emotion recognition. Pattern Recognition Letters., 120: 69-74.
[5] Raksarikorn, T., Kangkachit, T. (2018) Facial Expression Classification using Deep Extreme Inception Networks. In: 2018 15th International Joint Conference on Computer Science and Software Engineering. Nanjing, China. pp. 1-5.
[6] Zhang, S., Huang, T., Gao, W., Tian, Q. (2017) Learning affective features with a hybrid deep model for audio–visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology., 28(10): 3030-3043.
[7] Chollet, F. (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, Hawaii. pp. 1251-1258.
[8] Huang, Y., Jiang, Y., Hasan, T., Jiang, Q., Li, C. (2018) A topic BiLSTM model for sentiment classification. In: Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence. Shanghai, China. pp. 143-147.
[9] Gao, K., Xu, H., Gao, C., Hao, H., Deng, J., Sun, X. (2018) Attention-Based BiLSTM Network with Lexical Feature for Emotion Classification. In: 2018 International Joint Conference on Neural Networks. Rio de Janeiro, Brazil. pp. 1-2.
[10] Qureshi, A. H., Nakamura, Y., Yoshikawa, Y., Ishiguro, H. (2016) Robot gains social intelligence through multimodal deep reinforcement learning. In: 2016 IEEE-RAS 16th International Conference on Humanoid Robots. Mexico. pp. 745-751.