Multimodal data fusion algorithm applied to robots

Xin Zhang\textsuperscript{1,2}, Zhiquan Feng\textsuperscript{1,2,*}, Jinglan Tian\textsuperscript{1,2} and Xiaohui Yang\textsuperscript{1,2}

\textsuperscript{1}College of Information Science and Engineering, University of Jinan, Jinan, Shandong, 250022, China

\textsuperscript{2}Shandong 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, the use of multimodal human-computer interaction technology to achieve the enhancement of human intelligence has become a new topic in human-computer interaction research. When the robot can’t react correctly in a single mode, it is necessary to realize multimodal fusion. To this end, this paper proposes a multimodal fusion algorithm that applies the data obtained by the CNN feature layer to the decision-level. The speech recognition text is semantically matched with the text in the text library, and the similar probability vector is returned. At the same time, the similarity probability vector of the gesture recognition is obtained, and the data is filtered by the threshold, and the set of high probability data codes is assigned to the two modes. The intersection operation, and the final instruction is sent to the robot. The experimental results show that the influence of environmental factors on the single channel result is reduced, and the single mode ambiguity problem is eliminated. The multi-channel fusion algorithm with additional weight is more accurate than the common multi-channel fusion algorithm. At the same time, it has also been well received by many test users.

1. Introduction
Multimodal fusion usually includes multimodal fusion in the same modal information. Multimodal data can obtain more comprehensive and accurate information and enhance the reliability and fault tolerance of the system\cite{1}. This paper will focus on the multimodal information perception and fusion of robots, especially the work related to the fusion perception of vision and language. However, for human-computer interaction system, it needs strong real-time performance, and the environment is uncertain, so it needs a multimodal fusion algorithm with short time and high efficiency. These problems bring great challenges to the multimodal fusion perception of robot.

Therefore, this paper provides a decision-level multimodal human-computer interaction fusion algorithm based on additional threshold and weight ratio. The main advantages of this paper lie in that it not only solves the real time of human-computer interaction, but also solves the input problem of robot's hearing and vision in the complex environment. The main reasons are as follows: on the one hand, after increasing the threshold, more redundant data will be erased, which improves the speed of operation; On the other hand, after increasing the weight ratio, the influence of the environment on the two modes will be coordinated to make the fusion result more accurate.
2. Related work

2.1. Speech and gesture recognition
From 1959, American scholar B. Shackel started from the ergonomics of computer console design [2], which is considered to be the first document of human-machine interface. Humans already have the idea of interacting with machines. The way of human-computer interaction is mainly speech and posture. Speech recognition technology companies have developed speech recognition tools with high accuracy; gesture recognition was based on data gloves in the early days[5]. In 1983, Grime et al first used gloves with node markers to recognize gestures with palm bones, and completed simple cognition. In the 1990s, due to the advantages of accurate positioning of peripheral devices, Takahashi et al [3,5] Using data gloves to achieve the recognition of 46 specific gestures; later, the Fingerprint method replaced the data glove, Lee et al [4,5] used the information entropy algorithm to segment the background image and successfully applied it through the parallel computing algorithm. In the video data stream, the target image with an accuracy of 95% is identified.

2.2. Multimodal fusion
Later, because of the limitations of single-modal interaction, experts proposed multi-modal human-computer interaction. In [6], dense multimodal fusion (dmf) was realized by greedy stacking between different networks of different forms; later [7] Outlines the integration of gestures and speech recognition run results for storytelling applications; as the popularity of deep learning continues to grow, in [8,10], audio and video are implemented using hidden Markov models (chmm) Feature fusion; in addition, a dynamic Bayesian network is introduced in [9,10] for the fusion of graphics and speech features.

In summary, the above has carried out a lot of work in the multimodal fusion of feature level, and this paper mainly uses the data obtained by the feature layer to apply to the multimodal fusion based on decision-level, adds the weight value and threshold. the fusion is performed by the intersection operation, and finally the result instruction is transmitted to the robot.

3. Algorithm design
The multimodal human-computer interaction proposed in this paper includes three parts: information extraction, information processing and information output. the information acquisition module receives information from voice and gesture channels through input devices such as microphones and cameras, and then uses multimodal. The state information analysis fusion module generates multimodal cooperative expression content and synchronously outputs to the robot expression module.

Figure 1. Flow chart of multimode fusion algorithm.
3.1. Gestural intention and vocal intention inference
In this paper, a gesture library is established according to the daily gestures needed to complete relevant instructions. After the recognition by CNN, the similarity vector of the test gesture and other gestures in the recognition result is returned.

The text library is set up according to the instruction. The speech is recognized by CNN, the recognized text is matched with the text library semantically, and the similarity vector is returned.

Table 1. User Instruction Intent Decision Table.
| Gesture /text code | A: Say hello to the users | B: Give the user a glass of water and send it to the users | C: Bow to users | D: Turn on the TV | E: Shake hands with the users |

3.2. Multimodal fusion algorithm
After obtaining the similarity vectors of the two modes, it is necessary to carry out probability functionalization of the vectors of the two modes respectively, assuming that there is an array e, the ith element represented by ei, then the value Si of this element is:

\[
S_i = \sum_j e_j
\]

Table 2. User Instruction Intent Decision Table.

| MULTIMODAL FUSION ALGORITHM |
|-------------------------------|
| INPU: Audio, gesture          |
| T1: Robot execution           |
| OUT: Speech recognition       |
| PUT: Gesture recognition      |
| 1: Gesture recognition is carried out and similarity vector is obtained. Probability vector is obtained through probability function |
| 2: Delete the items less than 0.1 in the probability vector, assign the codes of the remaining items to the current mode as a set, and set the two modes as Audio and Gestures. |
| 3: Audio.intersection(Gestures)=Result[] |
| While(Length(Result)) |
| { |
| 4: If Length(Result)=1, Return Result[0]; |
| 5: Else Length(Result)>1, However, the results must be unique, and the two modalities RA,RG(RA+RG=1) are given different weights according to the actual situation of the experiment. Represents the weight values of speech and gestures, |
| If RA*P(Audio[A])+RG*P(Gestures[A])>RA*P(Audio[A])+RG*P(Gestures[A]) |
| Return A; |
| Else Return D; |
| } |
| While(Length(Result)==0) |
| { Find the text code and gesture code corresponding to Max(P(Audio[])) and |
Max(P(Gestures[])), find the text library and the gesture library to return the instruction content, use the voice to ask questions, and ask the user whether it is one of two instructions,

(1) The user answers that the robot asks the user to repeat the instruction content and find the text library or the gesture library return code;

(2) The user replies no, the program returns to Input.

Enter the code into the robot expression module and the robot responds.

7: End.

4. Experimental analysis and effect evaluation

The CPU selected at the host during the experimental operation is Intel(R) Core(TM) i7-4712MQ @ 2.30GHz, which uses an external ordinary microphone and camera to run under the 64-bit win10 system to control the robot in real time.

4.1. Demonstration of experimental results

Figure 2 shows the process from the actual experimental operation of the user, and we selected some representative points in the experiment to illustrate. The left 1 is the robot to map the user; the left 2 is the robot and the user handshake diagram; the right 2 is the robot and the user greeting diagram; the right 1 is the robot to help the user to open the TV map.

Figure 2. Experimental rendering.

4.2. Contrast experiment

In order to further test whether the multimodal fusion algorithm proposed in this paper meets the requirements, we first find 20 volunteers, using two-channel multimodal fusion and single-voice channel for experiment comparison in noisy environment; using dual channel in strong illumination environment. Multimodal fusion and experimental comparison of single gesture recognition channels. Then, the probability of success in 50, 100, and 150 comparison experiments in two environments was obtained. The comparison results are shown in Figure 3.

![Figure 3. Comparison chart of experimental accuracy.](image)

Figure 3. Comparison chart of experimental accuracy.

Figure 4 shows the influence of the weight value distribution on the experimental results during the experiment, in which the X-axis is the weight value of the gesture, and the Y-axis is the experimental environment in the laboratory (no simultaneous noise and glare environment factors) Success rate. In the ordinary decision-level multimodal fusion algorithm, speech and gesture each account for 50%, so it is also compared with the common decision-level multimodal fusion algorithm.
4.3. The user evaluation

In order to detect whether the fusion algorithm meets the design requirements and advantages and disadvantages, Figure 5 gives a comparison of NASA evaluations on the user's cognitive load during the experiment.

The user evaluation indicators are divided into mental demand, MD, physical demand, PD, performance, P, effort, E. And frustration, F. The NASA evaluation index uses a 5-point scale. Each indicator is divided into 5 levels.

5. Conclusion

In this paper, a single mode is easily affected by environmental factors, and it is easy to produce ambiguity problems. A method for applying the data obtained by the CNN feature layer to the decision-level multimodal fusion algorithm is constructed, which reduces the environmental factors. The impact of a single channel result received praise from volunteers. At the same time, there are some areas for improvement in the algorithm of this paper. On the one hand, the modality of fusion is still relatively small, so the accuracy will not reach the peak value; on the other hand, the types of instruction libraries are not enough, and the completed services are not enough. These are the key points for us to study and study in the future.

Acknowledgments

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

References

[1] Liu, H.P. (2017) Multimodal fusion sensing technology for robot. Chinese academy of command and control communication., 2.
[2] Dong, S.H., Wang, H. (2004) The human-computer interaction. Peking University press, Beijing.
[3] Takahashi, T., Kishino, F. (1991)Hand gesture coding based on experiments using a hand gesture interface device. ACM Sigchi Bull., 23: 67–74.
[4] Lee, J., Lee, Y., Lee, E., Hong, S. (2004) Hand region extraction and gesture recognition from
video stream with complex background through entropy analysis. In: The 26th Annual International Conference of the IEEE. San Francisco. pp. 1513-1516.

[5] Fang, W., Ding, Y., Zhang, F., Sheng, J. (2019). Gesture Recognition Based on CNN and DCGAN for Calculation and Text Output. IEEE Access, 7, 28230-28237.

[6] Hu, D., Wang, C., Nie, F., Li, X. (2019) Dense Multimodal Fusion for Hierarchically Joint Representation. In: Speech and Signal Processing. Barcelona. pp. 3941-3945.

[7] Kanawade, A., Varvadekar, S., Kalbande, D.R., Desai, P. (2018) Gesture and Voice Recognition in Story Telling Application. In: International Conference on Smart City and Emerging Technology. Tirunelveli. pp. 1-5.

[8] Yashwanth, H., Mahendrakar, H., David, S. (2004) Automatic speech recognition using audio visual cues. In: First India Annual Conference. Kharagpur. pp. 166-169.

[9] Liu, X., Zhao, Y., Pi, X., Liang, L., Nefian, A.V. (2002) Audio-visual continuous speech recognition using a coupled hidden Markov model. In: Seventh International Conference on Spoken Language Processing. Denver.

[10] Argyropoulos, S., Tzovaras, D., Strintzis, M.G. (2007) Multimodal fusion for cued speech language recognition. In: 15th European Signal Processing Conference. Poznan. pp. 1289-1293.