Multimodal Fusion of Speech and Gesture Recognition based on Deep Learning

Xiaoyu Qiu\textsuperscript{1,2}, Zhiquan Feng\textsuperscript{1,2}, Xiaohui Yang\textsuperscript{1,2} and Jinglan Tian\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. This paper proposes a multimodal fusion architecture based on deep learning. The architecture consists of two forms: speech command and hand gesture. First, the speech and gesture commands input by users are recognized by CNN for speech command recognition and LSTM for hand gesture recognition respectively. Secondly, the obtained results are searched by keywords and compared by similarity degree to obtain recognition results. Finally, the two results are fused to output the final instructions. Experiments show that the proposed multi-mode fusion model is superior to the single-mode fusion model.

1. Introduction

In previous studies, traditional machine learning models often control the robot through a single mode, but sometimes a mode is not enough to fully represent the recognized content\cite{1}. Therefore, the control method of multi-mode robot is introduced. Multi-mode representation learning is to eliminate the redundancy between modes by utilizing the complementarity between modes and improve the recognition rate \cite{2}. This paper proposes a multi-modal fusion architecture based on microphone-based speech command recognition and hand motion recognition based on Leap motion \cite{3}. In order to design the proposed architecture, this paper uses different deep neural networks to process the instructions of each single-mode input, and then detects the correctness of the recognition command through keyword retrieval and similarity comparison. Finally, the obtained results are merged and output instructions. The innovation of this paper is to compare the obtained recognition results twice and then to fuse. Firstly, the keywords of gesture recognition are retrieved in speech recognition. Secondly, the speech recognition words were processed by word segmentation, and then the word segmentation results were compared with the keywords of gesture recognition. Finally, the two comparison results were fused to obtain the final command, which improved the accuracy of identification.

2. Related work

2.1 Speech command recognition

In previous studies, speech commands have been used by many researchers in robot control. In the absence of a noisy environment, speech command recognition can achieve higher accuracy \cite{4}, but the robot in motion state inevitably generates its own noise. Different from ambient noise, the self-noise is closer to the microphone and more easily acquired. These factors greatly reduce the robustness of robot
speech recognition under motion [5]. CNN is superior to traditional rule-based feature extraction methods in terms of robustness and efficiency, and it has been successfully applied in a variety of applications, including speech command recognition [6]. Therefore, this study uses CNN for speech recognition.

2.2 Gesture Recognition
The initial gesture recognition mainly uses machine equipment to directly detect the angle and spatial position of the joints of the arms and arms [7]. Later, the vision-based gesture recognition technology makes the operation of the machine no longer requires physical contact between the operator and the machine [8]. With the increasing demand for application scenes, gesture recognition based on monocular cameras has received more attention. However, due to the large deformation of the gesture, self-occlusion, complex background [9], relying solely on gesture recognition can not meet the requirements of fast and accurate recognition by users. In hand motion recognition, the data set consists of a hand movement time series with strong order dependence between each hand movement [10]. Because LSTM has the advantage of learning long-term dependencies, LSTM is currently an ideal model for some sequence processing tasks.

2.3 Multimodal fusion
Multimodal models can potentially generate more knowledge about classification problems than single mode models. Multi-modal fusion in the HAR domain is usually divided into two categories: feature-level fusion and classification-level fusion. Feature-level fusion is the merging of feature data extracted from each modality before the classification process [11]. The purpose of classification-level fusion is to combine the classification results given by different modalities [12].

3. Algorithm design

3.1 Multimodal fusion architecture
The process of multi-modal fusion can be divided into multi-modal input and perception, identification and processing of multimodal information, integrate results to get instructions. Figure 1 shows the architecture of multimodal fusion. The words recognized by the speech command and the keywords recognized by the gesture can be represented as $X^S$ and $X^H$. $T_{iI}(X^S, X^H)$ and $T_{iD}(X^S, X^H)$ are the results obtained by using different alignment modes for the two modes.

![Figure 1. Multimodal Fusion Architecture Description.](image)

3.2 Algorithm Description
In the first process, the robot's microphone and camera respectively acquire the speech and gesture commands. In the second process, the speech recognition sentence and the gesture recognition keyword are first obtained. Secondly, the comparison is performed twice. The first comparison is to retrieve the keyword of the gesture recognition in the speech recognition sentence. Returns True if the retrieval is successful, otherwise returns False. The second comparison is to segment the speech recognition sentence, extract the keyword, and then compare the similarity with the keyword of the gesture recognition. If the similarity is above 60%, it returns True, otherwise it returns False. In the third process,
the results returned by the two sets of comparisons are intersected or combined, and when the processed result is True, the correct instruction is obtained. As shown in Table 1. And table 2 shows the actions corresponding to different instructions.

Table 1. Description of multi-mode fusion algorithm.

| Multimodal fusion algorithm |
|----------------------------|
| Input: Enter speech and gesture commands |
| Output: Output the recognized instruction |

1. (1) Identify and convert the input speech into text
   (2) While(Gesture number: i < The number of gestures in the recognition library: n)
     (1) Compare input gestures with the gestures of the library to get the similarity: P.
     (2) If the similarity of gesture: i > P, P = Pi.
     (3) Gesture number: i+1
   (3) Get the highest similarity gesture, and get the keyword
2. (1) Use jieba participle to process the text of the speech and obtain the keywords.
   (2) Use difflib's text comparison method to compare the keywords in the voice with the keywords of the gesture. If the similarity exceeds 60%, return True, otherwise return False.
3. if(Ti(XS, XH) ∩ TiD(XS, XH) = = True)
   { Instruction: comm = Gesture number: i }
else if(Ti(XS, XH) ∪ TiD(XS, XH) = = True)
   { (1) Compare the results of the two comparisons and take the highest accuracy as the recognition result.
     (2) Instruction: comm = Gesture number: i }
else:
   Instruction: comm = 0
End

Table 2. User Instruction Intent Decision Table.

| Output instruction | Corresponding behavior           |
|--------------------|--------------------------------|
| 1:                 | Take the phone for the user     |
| 2:                 | Take water for the user         |
| 3:                 | Respond to users               |
| 4:                 | Greet the audience             |
| 0:                 | Input command error, voice reminder retry |

4. Experiment

In this section, the author first describes the recognition process of each modality, and then shows the effect of multi-modal fusion, comparing the recognition rate of multi-modality and single-mode. Finally, the results of multimodal fusion are given.

4.1 Single mode identification

4.1.1 Speech Command Identification. This experiment uses the speech recognition data set provided by Baidu API. When the voice command is input, the sound is first framed. After the frame is divided, the
MFCC features are extracted to obtain a matrix. Finally, the matrix is transformed into text through three steps: the frame is recognized as a state, combine states into phonemes, combine factors into words. Figure 2 is an MFCC representation of recognized speech commands.

4.1.2 Gesture recognition. This experiment builds a gesture library based on the gestures required to complete the relevant instructions. When enters a gesture, the program controls the camera to take a photo and returns the gesture image to the temporary storage. In the recognition, the background will return the similarity between the image and the gesture in the library. If there is no gesture in the image, it returns 0. Finally, the gesture with the highest similarity is the recognition result, and the keyword is obtained. Figure 3 shows an example of a gesture corresponding to four sets of actions.

4.2 Multimodal fusion effect display
In the experiment, the author inputs speech and gesture commands to the robot. For example, when speaking to the robot, I want to make a call and make a call gesture. The robot will return the acquired voice in wav format and send the gesture image back. After the above single mode identification, the result is fused, and finally the result instruction is obtained and transmitted back to the robot. The experimental effect is shown in Figure 4.

4.3 Comparative experiment
In order to further detect whether the multimodal fusion achieved the expected accuracy, this experiment compared the results of multimodal fusion with the recognition results of single mode (speech, gesture). 15 experimenters were found to participate in the experiment. Each group performed 10 sets of actions, and then counted the number of successful recognitions. In order to be closer to daily life, this experiment was chosen to be carried out in a noisy and lit room, and the results as shown in Figure 5.
5. Conclusion
In this paper, a robust multi-modal fusion structure is proposed. In order to improve the accuracy of recognition, two single modes are first identified, and the recognition sentences and keywords are respectively obtained from the two commands of voice command and hand motion. The results of the two recognitions were further fused after two comparisons. Experiments show that the model has higher accuracy than the single-mode model.

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

References
[1] Liu, H., Wang, L. (2017) Gesture recognition for human-robot collaboration: A review. International Journal of Industrial Ergonomics., 68: 355–367.
[2] Liu, H., Fang, T., Zhou, T., et al. (2018) Towards Robust Human-Robot Collaborative Manufacturing: Multimodal Fusion[J]. IEEE Access., 6:74762-74771.
[3] Liu, H., Fang, T., Zhou, T., Wang, Y., Wang, L. (2018) Deep learning-based multimodal control interface for human-robot collaboration. Procedia CIRP., 72: 3–8.
[4] Wang, J. R., Gao, Y. C., Zhang, J. (2017) Kinect-assisted robot Speech recognition with noise [J]. Journal of tsinghua university: natural science., 09: 28-32.
[5] Zhang, H., McLoughlin, L., Song, Y. (2015) Robust sound event recognition using convolutional neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)., Brisbane, QLD. pp. 559–563.
[6] LeCun, Y., Huang, F. J., Bottou, L. (2004) Learning methods for generic object recognition with invariance to pose and lighting. In: IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)., Washington, DC, USA. pp. 97–104.
[7] Nweke, H.F., Ying, W. T., Al-Garadi, M.A., et al. (2018) Deep Learning Algorithms for Human Activity Recognition using Mobile and Wearable Sensor Networks: State of the Art and Research Challenges[J]. Expert Systems with Applications., 105: 10.
[8] Narain, B.R., Priscilla, D., Maria, S. (2017) Gesture Command Recognition System for Human Machine Interaction[J]. Australian Journal of Basic and Applied Sciences., 10: 41-45.
[9] Liao, L. L., Wang, Z. W., Wu, J. C., (2018) Gesture recognition based on monocular tracking-neural network classification algorithm [J]. Computer knowledge and technology., 14: 227-230.
[10] Zhu, G., Liang, Z., Shen, P., et al. (2017) Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM[J]. IEEE Access., 5: 4517-4524.
[11] Vaizman, Y., Ellis, K., Lanckriet, G., Weibel, N. (2018) ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems., New York. pp.554.

[12] Li, F., Shirahama, K., Nisar, M.A., Köping, L., Grzegorzek, M. (2018) Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors., Sensors. 18(3): 679-.