Hand Gesture and Speech Switchable Text Input Method for Wearable Augmented/Virtual Reality Devices

Ziyao Cheng, Qiang Chu, Gang Li, Chao Ping Chen*

Corresponding author's email: ccp@sjtu.edu.cn
Smart Display Lab, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

Keywords: augmented/virtual reality, text input method, hand gesture recognition, speech recognition.

ABSTRACT

We propose a hybrid text input method that caters to wearable augmented/virtual reality devices. When traditional input peripherals, e.g. mice and keyboards, are not available, this method can be an alternative by enabling users to type letters, numbers and symbols via either hand gesture or speech.

1 Introduction

Augmented/virtual reality (AR/VR) is preferably implemented on wearable devices [1–3]. Unlike unwearable devices, wearable devices are no longer equipped with the conventional input devices, *e.g.* keyboards, mice, touch panels *etc*. As a viable option, speech/voice recognition [2] has been widely adopted. However, speech/voice recognition suffers from several limitations. First, under noisy environment, speech recognition is not reliable. Second, for private or sensitive information, *e.g.* password, it is definitely not acceptable. Third, users have to passively accept the as-recognized texts. Fourth, for those who cannot talk, it is not applicable. When it comes to the above scenarios, hand gesture recognition could serve as a backup solution [3]. Motivated from the said issues, we present a text input method, which can be switched between the hand gesture and speech.

2 Proposed Method

2.1 Framework

Fig. 1 outlines the framework of the proposed text input method, which can be run in parallel between the hand gesture recognition module and speech recognition module. The pipeline of hand gesture recognition mainly consists of the hand detection, fingertip tracking, hand gesture command/control, handwriting recording, optical character recognition *etc*. The pipeline of speech recognition, on the other hand, mainly consists of wake-up words, voice command/control, speech recognition *etc*.

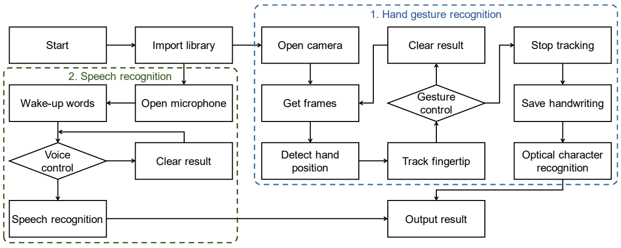


Fig. 1 Framework of the proposed text input method

2.2 Hand Gesture Recognition

Fig. 2 plots the network structure, which leverages the depthwise separable convolutions [4], for the hand gesture recognition. The network starts with an input of segmented hand of 224×224 pixels. The number of depthwise separable convolutional layer is 5. The output is classified into 4 different hand gestures, *i.e.* write/track, delete/clear, stop and save. The training and testing datasets contain 300 and 50 images, respectively. In our adjustment, the batch size is 10, number of epochs is 10 and learning rate is 0.001. The average accuracy could be 98.7% or above.

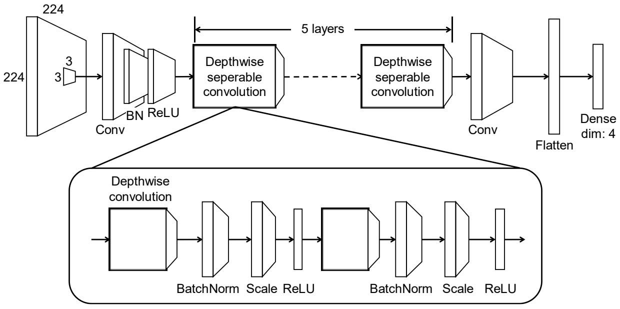


Fig. 2 Network structure for the hand gesture recognition

2.3 Speech Recognition

An end-to-end automatic speech recognition (ASR) [5], as shown in Fig. 3, is adopted. The core of ASR is essentially a recurrent neural network (RNN), which consists of invariant convolutional layers, gated recurrent units or bidirectional recurrent layers, and fully connected layers, in tandem with a connectionist temporal classification (CTC) layer [6].

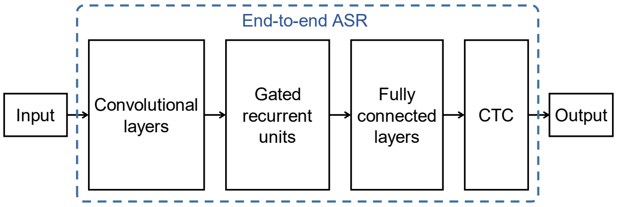


Fig. 3 Network structure for the end-to-end automatic speech recognition

3 Results and Discussion

To implement our text input method, an application is developed on the platform of Android-based devices. When operated in the hand gesture mode, as shown in Fig.

4(a), user needs to place his/her hand within the camera view. After the hand is detected (Fig. 4(b)), the continuous motion of fingertip as a whole will be recorded as one handwriting (Fig. 4(c)). In addition to the as-recognized text/letter, other candidates are also selectable to compensate the errors (Fig. 4(d)). When operated in the speech mode, as shown in Fig. 5, user shall use the wake-up words to trigger this mode. If the as-recognized texts are not correct, he/she can switch to the hand gesture mode so as to modify the texts.

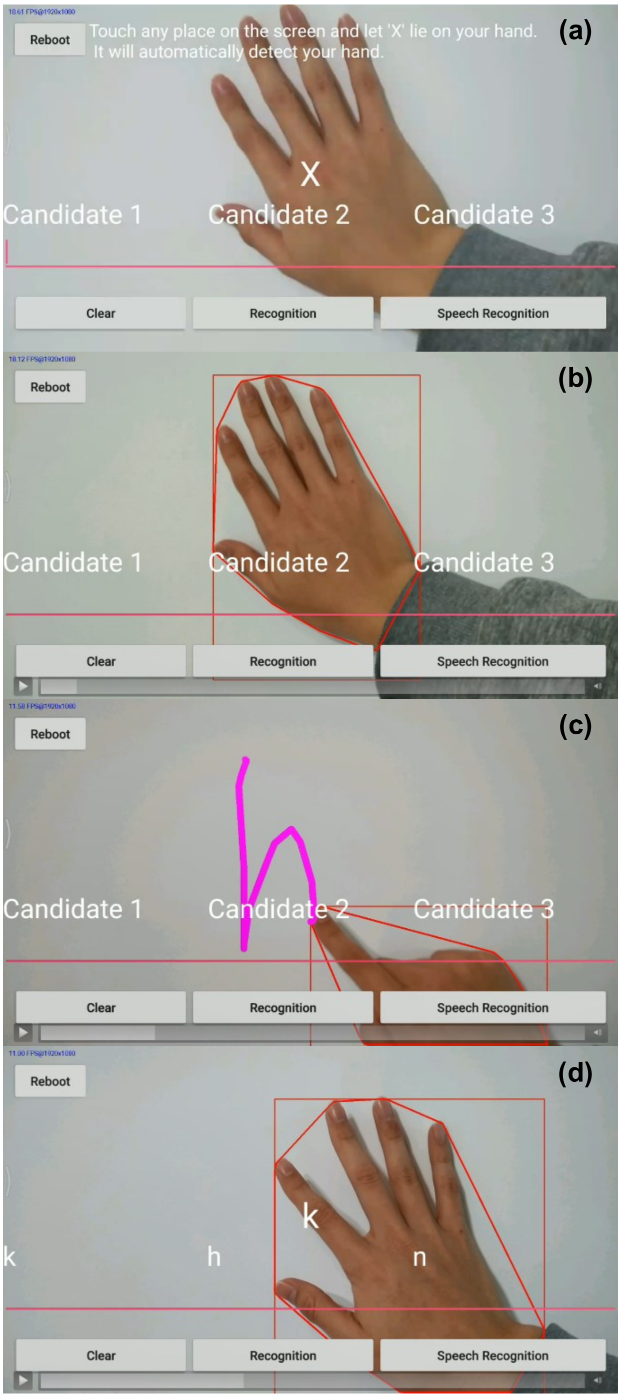
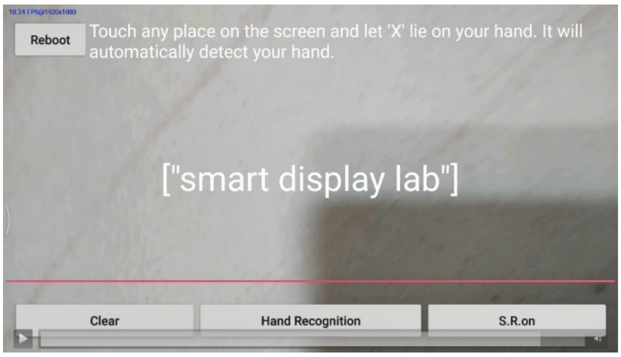

Touch any place on the screen and let 'X' lie on your hand. It will automatically detect your hand.

Fig. 4 Text input when in the hand gesture mode


Touch any place on the screen and let 'X' lie on your hand. It will automatically detect your hand.

["smart display lab"]

Fig. 5 Text input when in the speech mode (voice: smart display lab)

4 Conclusions

We have demonstrated a hand gesture and speech switchable text input method, which is intended to incorporate the merits of both techniques. To recognize the hand gesture, a depthwise separable convolutional neural network is configured. To recognize the speech, an end-to-end RNN-CTC model is adopted. In our experiments, an Android application has been developed.

5 Acknowledgments

This work is funded by National Natural Science Foundation of China (61901264, 61831015) and Science and Technology Commission of Shanghai Municipality (19ZR1427200).

References

[1] L. Mi, C. P. Chen, Y. Lu, W. Zhang, J. Chen, and N. Maitlo, “Design of lensless retinal scanning display with diffractive optical element,” Optics Express **27**(15), 20493–20507 (2019).

[2] J. Chen, L. Mi, C. P. Chen, H. Liu, J. Jiang, W. Zhang, “Design of foveated contact lens display for augmented reality,” Optics Express **27**(26), 38204–38219 (2019).

[3] C. P. Chen, L. Mi, W. Zhang, J. Ye, and G. Li, “Waveguide-based near-eye display with dual-channel exit pupil expander,” Displays **67**, 101998 (2021).

[4] J. W. Picone, “Signal modeling techniques in speech recognition,” Proceedings of the IEEE **81**(9), 1215–1247 (1993).

[5] S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” Artificial Intelligence Review **43**(1), 1–54 (2015).

[6] F. Chollet, “Xception: deep learning with depthwise separable convolutions,” 30th IEEE Conference on Computer Vision and Pattern Recognition, 1800–1807 (2017).

[7] D. Bandanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” 41st IEEE International Conference on Acoustics, Speech and Signal Processing, 4945–4949 (2016).