Speech + Posture: A Method for Interaction with Multiple and Large Interactive Displays

Xiaoxi Du[1], *Lesong Jia*[1], *Weiye Xiao*[1], *Xiaozhou Zhou*[1], *Mu Tong*[1], *Jinchun Wu*[1], *Chengqi Xue*[1,*]

[1] School of Machanical Engineering, Southeast University
Nanjing 211189, China

ABSTRACT

Multiple or Large displays play an important role in collaboration scenarios, because they can provide more display space. However, they are challenging concerning effective manipulating and managing the display contents, particularly when the displays are beyond the users' reachable region and operational limit region. In this work, we explore a particular interactive input combination for multiple or large displays, the Speech + Posture interactive input mode. We integrate postures to point to target display areas and phonetic keywords to designate display contents. This method makes interactive input commands concise and explicit, and it can support interaction with multiple or large interactive displays effectively.

Keywords: Speech, Posture, Input Modalities, Multiple Displays, Multimodal Interaction

INTRODUCTION

Tremendous changes brought by the rapid development of display technology are reflected in very specific real daily life. The use of larger displays and combinations of multiple screens is emerging in a growing number of application scenarios, such as integrated surveillance and commanding systems, multi-person seminars, indoor theme parks and multimedia teaching. Meanwhile, large displays and multi-screen combinations, which allow more content to be displayed, put new demands on interactive modes. The technically organic combination of the display and the interaction mode has the greatest impact on the design of an interactive system (Kim et al. 2018). How to interact with large or multiple displays that can display more content has attracted our attention.

RELATED WORK

Large and Multiple Displays. Large displays have larger screens and higher resolutions which can be viewed at greater distances in larger spaces (Kim et al. 2018), and they support split-screen, multiple pop-ups and picture-in-picture. Multiple displays can also present more information, while their combined use does not need to compromise the aspect ratio of the content being consumed (McGill et al. 2015). In addition, multiple displays can be spliced together into a large screen and display contents continuously.

Interactive Modes for Interactive Displays. For interactive displays, the main forms of interaction at present are :(1) using external devices; (2) interaction with physical objects; (3) multi-touch interaction; (4) voice interaction; (5) embodied and tangible interaction. Among them, voice interaction and embodied and tangible interaction are considered to be more effective interactive modes for large display systems, since voice interaction has the advantage of being natural, fast and hands-free, and embodied and tangible interaction using air gestures and body movements don't need to be near the screen or wear assistive devices (Ardito et al. 2015).

Multimodal Input. More and more scholars focus on making full use of the advantages of various interactive technologies to explore the multimodal human-computer interaction input combination. Some studies have combined eye tracking with gesture control to solve the problem of object manipulation in virtual space (Deng et al. 2017, Pfeuffer et al. 2017). Takeoka et al. (2010) designed an interactive table that combined air gestures and multi-touch. In addition, Pfeuffer et al.(2014) presented a technique based on the principle of "gaze selects, touch manipulates".

DESIGN AND IMPLEMENTATION

Consideration and Design

We try to migrate the natural behaviors of interpersonal communication to improve the naturalness and efficiency of human-computer interaction. In interpersonal interaction, humans often use verbal descriptions combined with gestures or postures to make the expression distinct, accurate and vivid. For example, when we spot a rabbit running on the lawn, we will raise our arms with our fingers pointed toward the running rabbit and tell our companions, "Look, a rabbit!". Combining our arm pointing and language description, the companions will quickly understand what we mean. On this basis, we propose the Speech + Posture technique for large and multiple interactive displays. As shown in Figure 1, the basic idea is to specify the display area through body posture, and then determine the specific display content through speech keywords. This technique can be used to solve the problem of manipulating and managing the display contents, particularly when the displays are beyond the users' reachable region and operational limit region.

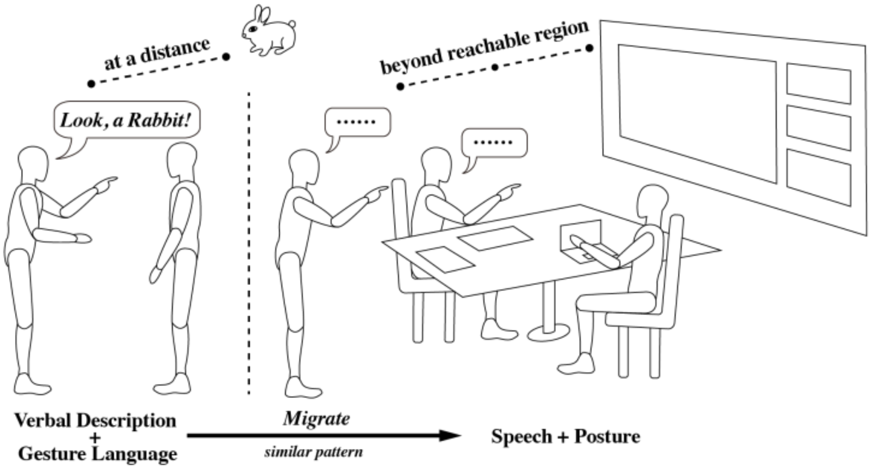

Figure 1. Design Concept of Speech + Posture Bi-modal Input

Implementation and Application

The recognition of gesture pointing direction and voice command is realized by Kinect and iFLYTEK open platform respectively. Then, a fusion algorithm based on logical operation will be executed to identify Speech + Posture instructions. To be more specific, when the two single-modal commands are recognized at the same time, the multi-modal fusion command is set off. For example, if the voice command A is recognized at the same time when the posture command A' appears, it can be determined that the multi-modal command "Speech A + Posture A'" is successfully recognized.

COMPARISON AND VALIDATION

Participants

20 right-handed subjects (10 female) aged between 22-26 years (Mean = 24, SD = 1.52) were recruited from the local university. All subjects had normal motor and linguistic abilities, and they all had normal naked visual acuity or corrected visual acuity.

Apparatus

A large display spliced by 12 Liquid Crystal Display of standard size was fixed in the wall, and it was divided into three sub-screens: left, middle and right according to the display areas. It was in a comfortable viewing position but out of arms' reach. An Azure Kinect was used to capture postures. A microphone was used for voice input. A timer was used to record task completion time. The apparatus was placed as shown in Figure 2.

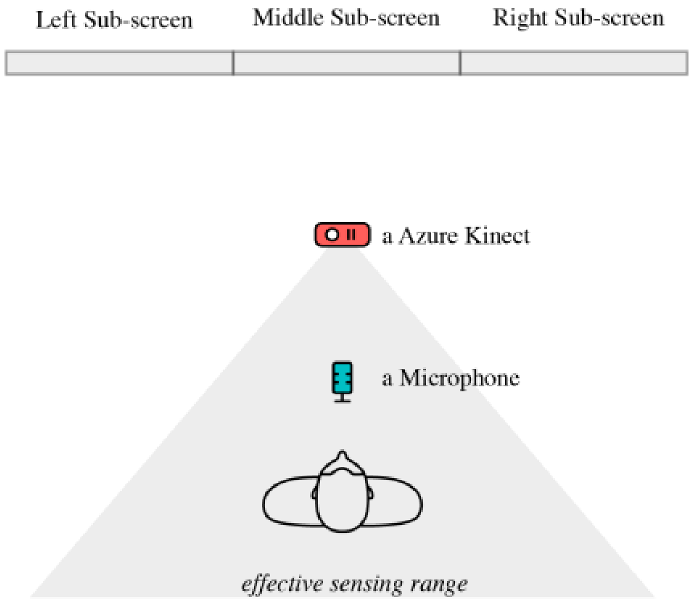

Figure 2. Position of the Apparatus

Procedure

Subjects were required to control the contents displayed in three display areas to change consecutively through the two interactive input modes: (i)use speech single-modal input, and the voice commands are detailed and explicit verbal descriptions; (ii)use phonetic keywords related to display contents and lift the left arm pointing to the target display area at the same time. A complete trial consisted of 13 consecutive subtasks and needed to be complete in each interactive modality. The task completion time of each trial was recorded.

Subjects were required to complete each subtask according to the instructions displayed on the sub-screen, and the 13 instructions appeared in random order. When the experiment began, a random instruction 1 appeared on the center sub-screen. After instruction 1 was operated, the random instruction 2 was displayed on the sub-screen working on currently as feedback, then the subjects could start on the next operation according to instruction 2. Subjects were asked to operate continuously until the last instruction was completed. Finally, Borg's Rating of Perceived Exertion Scale (REP) (Borg 1970) was completed to measure the subjective fatigue in the two interactive input modalities. The whole experiment process is shown in Figure 3.

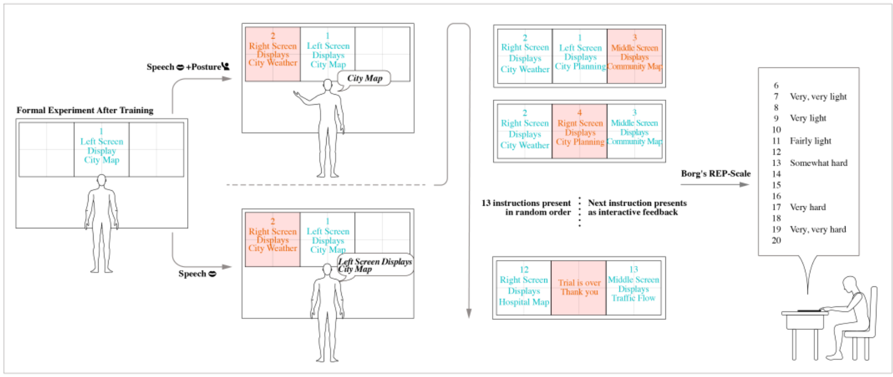

Figure 3. Experimental Procedure

Data Analysis

Task Completion Time Analysis. The task completion time of the Speech group was 86.70s±10.42, and the Speech + Posture group was 71.05s±10.62. The average completion time of the combined input group was 15.65s faster than that of the Speech group. There was no significant difference in the standard deviation of the task completion time between the two groups, but the coefficient of variation of Speech input (12.02%) was smaller than that of combined speech and posture input (14.95%). According to the boxplot graph of task completion time (see Figure 4a), the data distribution of the Speech + Posture group presented positive skewness: mean (71.05s) > median (69.5s) > mode (66s). T-test analysis was performed after eliminating data outliers, and the result was T=8.58, P<0.05(α=0.05), which meant that the difference between the two groups had statistical significance.

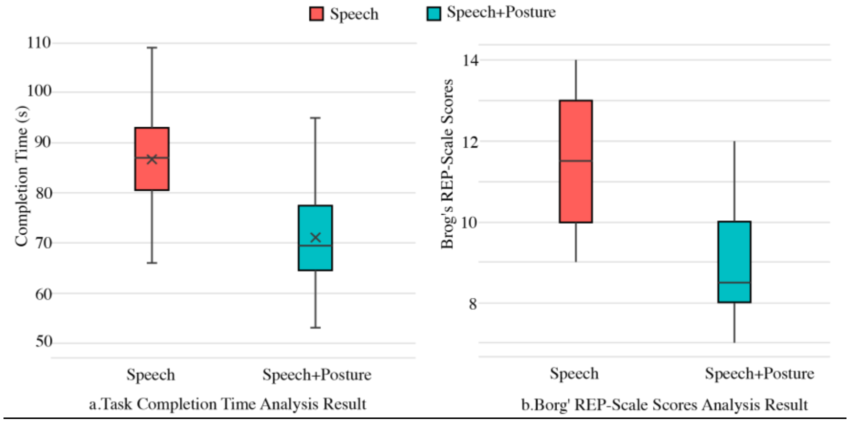

Figure 4. Task completion time and Brog’s REP-Scale scores analysis results

Borg’s Rating of Perceived Exertion Scale(REP) Analysis. The Borg’s REP-Scale scores of 20 subjects for the two interaction modalities are shown in Figure 4b. The score of Speech input was between 10 and 13, and the score of Speech + Posture input was between 8 and 10. Compared with speech single-modal input, the bi-modal input of combining speech and posture had significantly lower energy and cognition consumption.

DISCUSSION

The results of task completion time show that the time required for Speech + Posture input was shorter when completing tasks with the same difficulty. Compared with Speech input, the interaction efficiency was improved by 18.05%. The variation coefficient of task completion time in the Speech + Posture group was relatively large, and it indicated that the subjects might be slightly less familiar with the bi-modal input compared with the Speech input, resulting in a slightly larger standard deviation of task performance. The data of the Speech + Posture input group were also more dispersed, which indicated that the interactive performance of the group was unstable. However, according to the extremum data of the Speech + Posture group, some subjects could complete the continuous sub-task of specifying contents within 55s at the soonest. Moreover, the boxplot of task completion time is positively skewed, indicating that task performance has the potential to improve. If users' proficiency can be improved, interaction efficiency can be further improved.

Speech + Posture input is superior to Speech input in terms of alleviating cognitive frustration. The possible reason is that the words of voice commands required for Speech single-modal input are too long including three key parts: object, action and content. The error of any part would affect the operation result. In the experiment, sometimes the subjects needed to repeat voice commands several times to be recognized, and this process took a number of cognitive resources. The Speech + Posture input can make "position pointing" and "content setting" execute at the same time—the subjects use keywords related to contents as voice commands when the arm points to the sub-screen. The voice command "Left Screen Displays City Map" is simplified to "City Map". The fewer words and the higher recognition rate also greatly improved the task performance. Moreover, the novel interactive combination of Speech + Posture brought some freshness to the subjects and was considered to be in line with daily habits with high acceptance.

CONCLUSION

This study explored the feasibility of Speech + Posture bi-modal interactive input to control the display content of specified areas in large displays or multiple displays. We conclude that Speech + Posture input can take full advantage of position pointing of the arms and content description of the language, which greatly improves the efficiency and stability of human-computer interaction. Our work demonstrates that multimodal interaction design will be more reasonable and effective when it abides by the human innate interaction idioms. In the future, we will explore the suitability between interactive technologies and interactive tasks, in order to give full play to the interaction advantages of different modalities to achieve more natural interaction.

ACKNOWLEDGMENTS

This work was supported jointly by National Natural Science Foundation of China (No. 71901061).

REFERENCES

Ardito, C., Buono, P., Costabile, M. F., Desolda, G. (2015). “Interaction with large displays: A survey.” *ACM Computing Surveys (CSUR)*, 47(3), pp. 1-38.

Borg, G. . (1970). “Perceived exertion as an indicator of somatic stress.” S*candinavian journal of rehabilitation medicine*, pp. 92-98.

Deng, S., Jiang, N., Chang, J., Guo, S., & Zhang, J. J. (2017). “Understanding the impact of multimodal interaction using gaze informed mid-air gesture control in 3D virtual objects manipulation.” *International Journal of Human-Computer Studies*, 105, pp. 68-80.

Kim, D., Kim, H. M., Kim, H. K., Park, S. R., Lee, K. S., & Kim, K. H. (2018). “ThunderPunch: A bare-hand, gesture-based, large interactive display interface with upper-body-part detection in a top view.” *IEEE computer graphics and applications*, 38(5), pp. 100-111.

McGill, M., Williamson, J. H., Brewster, S. A. (2015). “A review of collocated multi-user TV.” *Personal and Ubiquitous Computing*, 19(5), pp. 743-759.

Pfeuffer, K., Mayer, B., Mardanbegi, D., & Gellersen, H. (2017). “Gaze+ pinch interaction in virtual reality.” *In Proceedings of the 5th Symposium on Spatial User Interaction*, pp. 99-108.

Pfeuffer, K., Alexander, J., Chong, M. K., & Gellersen, H. (2014). “Gaze-touch: combining gaze with multi-touch for interaction on the same surface.” *In Proceedings of the 27th annual ACM symposium on User interface software and technology*, pp. 509-518.

Takeoka, Y., Miyaki, T., & Rekimoto, J. (2010). “Z-touch: an infrastructure for 3d gesture interaction in the proximity of tabletop surfaces.” *In ACM International Conference on Interactive Tabletops and Surfaces*, pp. 91-94.