Multi-modal human-computer interaction system in cockpit

Jie Ren¹,a, Yanyan Cui¹, Jing Chen¹*, Yuanyuan Qiao¹ and Luhui Wang²

¹ School of Optics and Photonics, Beijing Institute of Technology, Beijing, 100081, China
² School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
*Corresponding author’s e-mail: chen74jing29@bit.edu.cn
a3120180562@bit.edu.cn

Abstract. In order to explore new human-machine interaction methods, a set of multi-modal human-machine interaction coordinated control system is proposed, which realizes the basic flight control based on the change of the pilot's field of view, touch control, voice control and other information obtained from multi-mode coordinated control. This system introduces a new type of human-computer interaction into the cockpit application on the basis of the human-computer interaction interface for flight driving, and has carried out research on multi-mode collaborative control system, mainly including eye movement interaction, touch interaction gesture Interaction and voice interaction. Finally, this project formed a multi-mode cooperative control software and hardware system.

1. Introduction
With the gradual development of the modern aviation industry, pilots need to pay attention to the mode control panel, main flight display, navigation display, flight management system and other panels at the same time. Therefore, it requires a new interactive method to reduce the pilot's work pressure and ensure the accuracy and safety of the flight. In terms of interaction methods, people usually instinctively use multiple methods to interact at the same time [1-2]. Multi-channel human-computer interaction can comprehensively utilize the advantages of different interaction methods and choose the most suitable interaction method according to different scenarios [3-4]. At present, there is a series of multi-modal human-computer interaction research [5-8], especially in the vehicle and airborne environment [9-10].

This paper constructs a set of multi-mode cooperative control system according to the pilot's field of view changes, touch control and other information, designs a flight pilot human-machine interface based on the decomposition of the flight task, realizes the multi-mode cooperative control flight driving operation and explores scenarios and tasks suitable for each channel.

2. Task analysis and single channel design
Taking take-off, cruise, and landing as an example, the flight task is divided into sub-tasks. A specific task is decomposed and expressed as a series of sub-tasks $T = \{t_1, t_2, t_3, ..., t_n\}$. From the perspective of future development trends, the multi-modal human-computer interaction mode is the most natural and ideal human-computer interaction with the least cost of learning. The multi-channel human-computer interaction studied in this subject is divided into four channels, touch, gesture, eye movement and voice.
2.1. Flight Task partitioning

In order to analyse the flight tasks, the tasks of the whole process are abstracted into five types of tasks: Click, Rotation, Input, Zoom and Sliding. The tasks in the take-off phase, cruise phase, and landing phase are specifically divided as follows:

| Serial number | Flight phase | Name                                             | Task type | Serial number | Flight phase | Name                                             | Task type |
|---------------|--------------|--------------------------------------------------|-----------|---------------|--------------|--------------------------------------------------|-----------|
| 1             | take-off     | BrakesToggleRegular                              | Click     | 11            | landing      | SpeedBrakesDownFull                              | Rotation  |
| 2             |              | AutopilotServosToggle                            | Click     | 12            |              | FlapsDownFull                                    | Rotation  |
| 3             |              | FlapsDown                                        | Rotation  | 13            |              | EnginesThrottleDown                              | Rotation  |
| 4             |              | EnginesThrottleUp                                | Rotation  | 14            |              | EngineThrustReserve                              | Click     |
| 5             | cruise       | LongitudeInput                                   | Input     | 15            |              | EngineThrustReserveFull                          | Rotation  |
| 6             |              | LatitudeInput                                    | Input     | 16            |              | BrakesToggle                                     | Click     |
| 7             |              | ZoomOutTheMap                                    | Zoom      | 17            |              | CloseEngineReserve                               | Click     |
| 8             |              | ZoomInTheMap                                     | Zoom      | 18            |              | CloseTheEngine                                   | Rotation  |
| 9             |              | LeftSlideTheMap                                  | Sliding   | 19            |              | CloseTheSpeedBrakes                              | Rotation  |
| 10            |              | RightSlideTheMap                                 | Sliding   | 20            |              | CloseTheFlaps                                    | Rotation  |

2.2. Touch interaction

Touch is currently one of the most commonly used interaction methods. There are some topics to study the combination of touch and other channels\cite{11,12}. We use ViewSonic 27-inch touch screen as hardware.

Above we abstracted the flight process into five tasks, and now we need to design how each channel implements these five tasks. The Click task is to touch the screen with your finger, and the Rotation task is to rotate around the centre point with two fingers. The Input task is to input numbers or letters in sequence and to press confirm at the end. The Zoom task is to control the change of the distance between two fingers. The Sliding task is to slide left or right without stopping. Specific operations are shown in Table 2.

| Task                      | Click | Rotation | Input | Zoom | Sliding |
|---------------------------|-------|----------|-------|------|---------|
| Touch gesture             |       |          |       |      |         |

2.3. Gesture interaction

Leap Motion is a somatosensory controller with two high-definition cameras and three infrared lights. The dual cameras calculate gesture poses through gesture screens from different angles, and infrared lights compensate for lights. Leap Motion can collect the position and posture information of the palm, fingers, and knuckles, and capture the user's gesture motion data. These data can be used to determine the gesture category.

For Click tasks, gesture interaction can imitate mouse operations, mapping the position of the three-dimensional index fingertip to the two-dimensional screen, and a period of static means clicking the position. The task of Rotation is to first select the object by clicking, and then to swipe left or right with the index finger making the target to rotate left or right. The Input task is to click one by one, and to click confirm after entering all the information. In the Zoom task, we set up two states: Palm Extended and Fisted. When switching from Palm Extended to Fisted, the screen is reduced, and vice versa. The Sliding task is to straighten the index finger and middle finger, and to swipe left or right with two fingers.
Table 3 Gesture interaction for different tasks

| Task   | Click | Rotation | Input | Zoom   | Sliding |
|--------|-------|----------|-------|--------|---------|
| Gesture| ![Gesture Click](image1) | ![Gesture Rotation](image2) | ![Gesture Input](image3) | ![Gesture Zoom](image4) | ![Gesture Sliding](image5) |

2.4. Eye movement interaction

Tobii Pro desktop eye tracker which is specially designed for extensive human behaviour research uses the latest technology eye tracking system and is capable of capturing rapid eye movements. The eye tracking system has excellent accuracy, precision and tracking stability and has achieved wide compatibility with various test groups in the actual research environment.

After calibration, eye movements are divided into gaze and eye track. Click and Input tasks are triggered by gaze and the rest are triggered by eye track. The trajectories are clockwise, counter clockwise, left, right, bottom left, and top right.

Table 4 Eye movement interaction track

| Task   | Click | Rotation | Input | Zoom | Sliding |
|--------|-------|----------|-------|------|---------|
| Interaction track | ![Interaction Click](image6) | ![Interaction Rotation](image7) | ![Interaction Input](image8) | ![Interaction Zoom](image9) | ![Interaction Sliding](image10) |

2.5. Voice interaction

The voice acquisition hardware is the Edifier headset, which is an integration of a headset and a microphone. The voice interaction is based on MASR voice recognition. 30 students were invited to record 200 flight-related instructions. After training, the accuracy rate on the verification set reached 99.50%. The trigger text of each task is shown in Table 5.

Table 5 Voice interaction words for different tasks

| Task   | Click | Rotation | Input      | Zoom     | Sliding |
|--------|-------|----------|------------|----------|---------|
| Interactive words | "press" | "rotate" | number/letter | "zoom out" | "previous page" |

3. Multimodal human-computer interaction system

This topic will carry out the research of multi-mode cooperative control system, mainly including eye movement interaction, touch interaction, gesture interaction, voice interaction and propose multi-mode cooperative interaction optimization strategies for typical flight tasks.

Figure 1. Block diagram of a multi-mode collaborative system
3.1. Experimental parameters

3.1.1 Time parameter

We are going to test the performance of the four channels on five tasks. First of all, select a task on the initial page and complete the task without stopping as quickly as possible after the experiment starts. The tester need to keep their attention and starts to use a channel to complete the task. Step by step replace tasks and interactive channels and complete all experiments.

In the human-computer interaction, the important parameter can be clearly quantitative analysis is time, so four time points are defined and the specific definitions are shown in Figure 2.

Therefore, we define three times: response time, operation time, and system identification time.

\[ \text{response time} = \text{time}_1 - \text{time}_0 \]
\[ \text{operation time} = \text{time}_2 - \text{time}_1 \]
\[ \text{system identification time} = \text{time}_3 - \text{time}_1 \]

After recording all the above data, calculate the time and plot the data respectively, as shown in Figure 3. The figure is divided into five parts. Each part corresponds to a type of task and each curve is divided into three nodes, which are reaction time, operation time, and system identification time. The four curves represent the four interactive channels of eye movement, touch, gesture and voice.

3.1.2 Subjective evaluation

In addition to objective indicators, subjective evaluation is also required. The evaluation of comfort, fatigue, and satisfaction of the four modalities by testers was investigated through questionnaires. In order to unify the standard and facilitate the optimization, the evaluation result is in the range of [0,1] with 0 representing the best and 1 representing the worst. The result is shown in Figure 3 above.

3.2. Dynamic programming optimization

3.2.1 Energy equation

This project intends to establish an interaction mode energy equation based on the performance evaluation of a single interaction mode. The energy equation can calculate the summation evaluation on the six evaluation indicators designed and by adjusting different weights the best expectation for a certain performance can be obtained. The energy equation is written as:
\[ \text{Cost}_{\text{total}}(T) = \sum_{t \in T} \left[ \lambda_1 \text{Cost}_{\text{ref}}(t) + \lambda_2 \text{Cost}_{\text{ope}}(t) + \lambda_3 \text{Cost}_{\text{sys}}(t) + \lambda_4 \text{Cost}_{\text{com}}(t) + \lambda_5 \text{Cost}_{\text{fat}}(t) + \lambda_6 \text{Cost}_{\text{sat}}(t) \right] \]

\[ \text{Cost}_{\text{total}}(T) \] represents the total interactive cost value for a certain task sequence \( T = \{t_1, t_2, t_3, \ldots, t_N\} \), \( \text{Cost}_{\text{ref}}(t) \), \( \text{Cost}_{\text{ope}}(t) \), \( \text{Cost}_{\text{sys}}(t) \), \( \text{Cost}_{\text{com}}(t) \), \( \text{Cost}_{\text{fat}}(t) \), and \( \text{Cost}_{\text{sat}}(t) \) represent the six cost values of response time, operation time, system identification time, comfort, fatigue, and satisfaction, respectively; \( \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 \), and \( \lambda_6 \) represent their weights. Users can change the weight of the six evaluation indicators according to the expected demand of a specific task. For example, in a mission with high pilot accuracy, the weight of \( \lambda_1 \) can be increased accordingly.

### 3.2.2 Dynamic programming

Dynamic programming is an algorithm for the optimization of decision-making. It is suitable for dividing the process into several stages and each stage makes a decision to achieve the best effect of the whole process. Our flight mission decision analysis is also subdivided into subtasks, so dynamic programming algorithms are applied. The specific realization of the algorithm is to calculate the four modalities of the previous task and the four modalities of the current task one-to-one, and consider the switching time cost to select the best result. In addition to the performance of each mode itself, switching between different modes will also cause a burden on people. Therefore, after all data is standardized, the modal switching cost is quantified by measuring the switching time data between the two modes of adjacent tasks.

### 3.2.3 Switch time

When we use two modalities between two consecutive subtasks, we think that there is a certain time cost when people switch between different modalities. In order to reduce the complexity of the problem, we assume that the switching time cost is only related to the two modes before and after. 4 modes are combined into 16 switching modes and each switching mode is composed of two tasks before and after. The switching time cost table obtained after averaging multiple experiments is shown in Figure 4.

![Switching costs between different modes](image)

**Figure 4. Switching costs between different modes**

### 4. Experiment

#### 4.1 Experiment environment

In order to study the performance of different modes on different tasks, we conducted a multi-mode time test experiment. A total of 10 graduate students participated in the experiment. The touch screen is connected to the host via HDMI cable and USB cable to display the human-computer interaction interface. Another screen displays the X-Plane flight picture as shown in the Figure 5. Leap Motion is connected to the host via a USB cable and placed on the desktop with the front face up. At the same time, wear headphones and adjust the position of the microphone appropriately. The desktop eye tracker is placed at the bottom of the screen. The experimental environment is shown in Figure 5.
4.2. Flight simulation software X-Plane

Considering that it is difficult to construct a real cockpit environment, the X-Plane flight simulation software is used for experiments. In X-Plane, the flight status of the aircraft, detailed modern aircraft simulator, 3D cockpit, intuitive user interface and other aspects are fully simulated.

4.3. Experimental Results

When the weight parameters are uniformly normalized to 1/6 and the mission sequence during cruise is taken as an example, the task sequence is composed of input, input, zoom, zoom, slide, and slide. The best result of the task sequence is calculated and displayed in the table below. There are also some random combination methods with the calculated energy value. Through comparison, it can be seen that the dynamic programming algorithm is indeed the best result.

| Mission sequence during cruise | Energy   |
|-------------------------------|----------|
| input | input | zoom | zoom | slide | slide |         |
| Optimal sequence              |          |
| touch | touch | voice | voice | touch | touch | 2.3407  |
| Random sequence               |          |
| eye movement | touch | gesture | eye movement | touch | eye movement | 3.7214  |
| gesture | touch | touch | eye movement | gesture | gesture | 3.4407  |
| touch | voice | touch | voice | eye movement | voice | 3.0637  |
| voice | voice | gesture | eye movement | touch | eye movement | 3.5213  |
5. Conclusion
This topic has realized the multi-modal human-computer interaction method based on touch gesture/eye movement/voice and verified it in X-Plane. You can choose any interactive method for operation at any time. And in fact, different task stages usually have different target requirements. Under different expectations, the optimal combination of operation sequences can be calculated through quantification, so that the combination of tasks at different stages can be optimized.

References
[1] R. Zhao, K. Wang, R. Divekar, R. Rouhani, H. Su and Q. Ji, "An Immersive System with Multi-Modal Human-Computer Interaction," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi’an, 2018, pp. 517-5241427.
[2] Turk, M., Multimodal interaction: A review. Pattern Recognition Letters, 2014. 36: p. 189-195.
[3] Baig, Muhammad Zeeshan & Kavakli, Manolya. (2019). Connectivity Analysis of Functional Brain Networks in Using Multi-modal Human–Computer Interaction.
[4] Lu, L., Lyu, F., Tian, F. et al. An exploratory study of multimodal interaction modeling based on neural computation. Sci. China Inf. Sci. 59, 92106 (2016).
[5] Yang S., Pan Y. (2014) A Study on Methods of Multimodal Interaction in Vehicle Based on Wheel Gestures and Voices. In: Stephanidis C. (eds) HCI International 2014 - Posters’ Extended Abstracts. HCI 2014. Communications in Computer and Information Science, vol 434. Springer, Cham.
[6] Karpov, A.A., Yusupov, R.M. Multimodal Interfaces of Human–Computer Interaction. Her. Russ. Acad. Sci. 88, 67–74 (2018).
[7] Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen C. F. Chan. 2016. Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. SIGAPP Appl. Comput. Rev. 16, 3 (September 2016), 37–49.
[8] David Rozado, Jason Niu, and Martin Lochner. 2017. Fast Human–Computer Interaction by Combining Gaze Pointing and Face Gestures. ACM Trans. Access. Comput. 10, 3, Article 10 (August 2017), 18 pages.
[9] Ken Pfeuffer and Hans Gellersen. 2016. Gaze and Touch Interaction on Tablets. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST ’16). Association for Computing Machinery, New York, NY, USA, 301–311.
[10] Shi Y., Li X., Ouyang D., Jiang H. (2017) Multimodal Usability of Human-Computer Interaction Based on Task-Human-Computer-Centered Design. In: Huang DS., Bevilacqua V., Premaratne P., Gupta P. (eds) Intelligent Computing Theories and Application. ICIC 2017. Lecture Notes in Computer Science, vol 10361. Springer, Cham.
[11] Zongkai Tian and Yan Zhang. 2020. A Dempster-Shafer Evidence Theory Based Multimodal Human Gesture Recognition Method. In Proceedings of 2020 the 6th International Conference on Computing and Data Engineering (ICCDE 2020). Association for Computing Machinery, New York, NY, USA, 147–153.
[12] Abdul Rafey Aftab. 2019. Multimodal Driver Interaction with Gesture, Gaze and Speech. In 2019 International Conference on Multimodal Interaction (ICMI ’19). Association for Computing Machinery, New York, NY, USA, 487–492.