

Article

# Multimodal Fusion of Voice and Gesture Data for UAV Control

Xiaojia Xiang, Qin Tan \*, Han Zhou, Dengqing Tang and Jun Lai

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

\* Correspondence: tanqin20@nudt.edu.cn

**Abstract:** To enable unmanned aerial vehicle (UAV) operators to efficiently and intuitively convey their commands to a swarm of UAVs, we propose the use of natural and human-centric input modalities, such as voices and gestures. This paper addresses the fusion of input modalities such as voice and gesture data, which are captured through a microphone and a Leap Motion controller, respectively, to control UAV swarms. The obtained experimental results are presented, and the achieved performance (accuracy) is analyzed. Finally, combined human factor ergonomics test with a questionnaire to verify the method's validity.

**Keywords:** unmanned aerial vehicle (UAV) control; voice; gesture; multimodal fusion

## 1. Introduction

Advanced computer-controlled environments such as virtual reality (VR) and immersive displays are manifestations of efficient visual information for unmanned aerial vehicle (UAV) operators. These displays are designed to simplify interactions with complex information and they often present information in a human-centered manner. However, existing user-input methods, e.g., mice and keyboards, are no longer convenient for interacting with these immersive displays and remotely controlling UAV missions. To improve the effectiveness of communication and reduce the time and effort required to complete tasks, human-computer interaction must be as fluid as manipulating a "natural" environment.

In recent years, natural interactions have attracted considerable interest. Psychological studies have also shown that people prefer to use gestures in combination with speech in virtual environments because they are easy to learn for the operator [1]. Gesture-based natural interaction systems are the most intuitive system type. Gesture-recognition technology provides a simple, fast, and efficient human-computer interaction environment, allowing an operator to issue instructions to the system of interest through simple gestures. In addition, speech-based natural interaction systems provide better system control. Voice-recognition features allow users to communicate naturally with a system using spoken language. These interaction methods have significantly improved the efficiency and comfort of human-computer interactions. However, the use of a single interaction mode often results in low fault tolerance and few application scenarios. Gestures and speech together constitute language. They have a bidirectional and mandatory influence on each other, meaning that people usually consider both simultaneously [2]. Multimodal input has many benefits, especially when dealing with gesture and speech combinations. When gestures and speech are present simultaneously, it helps reduce faster task completion times and even lower error rates [3].

This paper aims to address the challenges faced by users when controlling UAVs in virtual environments using multiple natural interactions. This scenario combines multiple modes, allowing the system to more accurately and naturally perform predefined tasks. The design combines voice commands with captured gestures. The remainder of this article is organized as follows. In Section 2, the related work on multimodal systems is discussed. In Section 3, the multimodal interaction integration system is proposed, and two



**Citation:** Xiang, X.; Tan, Q.; Zhou, H.; Tang, D.; Lai, J. Multimodal Fusion of Voice and Gesture Data for UAV Control. *Drones* **2022**, *6*, 201. <https://doi.org/10.3390/drones6080201>

Academic Editor: Andrey V. Savkin

Received: 11 July 2022

Accepted: 9 August 2022

Published: 11 August 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

aspects, voice recognition and gesture interaction, are described for data fusion. The system experiments including accuracy test and human factor ergonomics test are presented in Section 4, and the conclusions are given in Section 5.

### *Contributions*

In this paper, we propose the use of natural and human-centric input modalities, such as voices and gestures. The main contributions are as follows:

- We propose a multimodal interaction integration system that provides intelligent interaction service support for command recognition by dynamically modifying the weights to fuse gesture and speech data to achieve the accurate output of UAV control commands in virtual environments.
- We investigated gesture-only and speech-only command systems to highlight the respective advantages of these input modes. Meanwhile, we studied the integrated multimodal interaction system with an accuracy of 95.8824% to emphasize the synergy found when combining these modalities. This result is higher than the unimodal method and better than other fusion methods of the same type.
- We constructed the UAV flight simulation for the study of UAV mission operations and combined it with a human-machine efficacy questionnaire to verify the method's validity.

## **2. Related Work**

Virtual-reality environments were used for an exploratory study to investigate multimodal interaction to control a swarm of UAVs [4]. Furthermore, previous work demonstrates that a multimodal approach can direct interaction with UAVs, for example, by taking off and landing through speech and controlling movement by gesture [5,6]. Multimodal interaction can address high information loads and communicate within various environmental constraints. Typical multimodal human-robot interaction is possible through two methods.

- Accepting operator inputs from separate devices.
- Accepting operator inputs as different modes and fusing the input to capture commands related to operator behaviors.

The first method, which switches between multiple input devices, increases the cognitive overhead, resulting in stress and performance losses [7]. Thus, this method is not suitable for complex UAV-based mission control environments. Therefore, the second method is the focus of this paper.

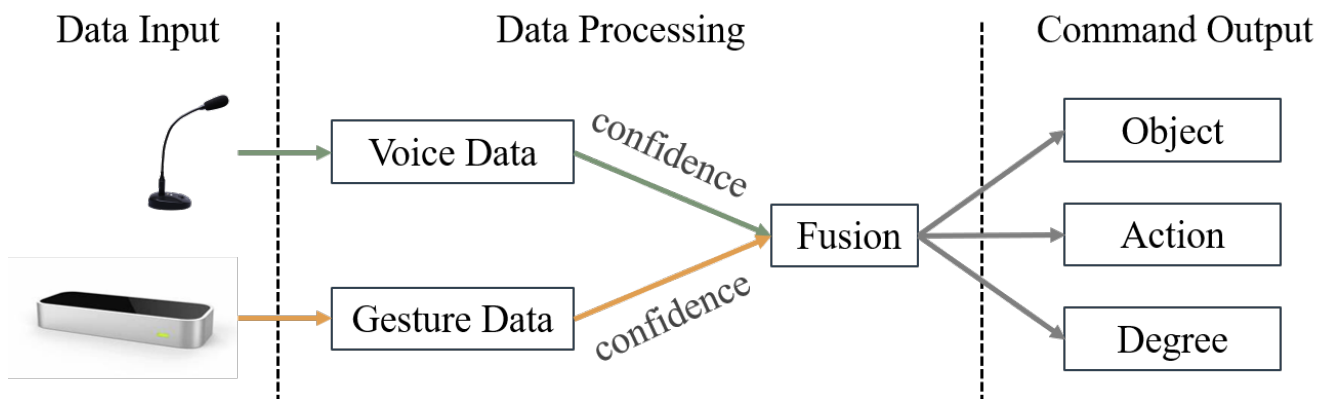
In the field of human-computer interaction, various methods use multimodal interactions to improve human-computer communication. However, many systems are characterized by setting the priority levels of their input modes. For example, in [8], Burger et al. proposed a probabilistic and multihypothesis interpretation framework for fusing speech and gesture component results. The primary input mode is speech. When the interpreter needs a supplementary gesture for disambiguation, the system fuses the gesture interpretation results provided in the same time window. In [9], Tauheed et al. fused voice data and captured electromyogram data with an MYO band based on priority. When the MYO band could not capture electromyogram signals, voice commands were used to compensate for this shortcoming and served as the only input, improving the accuracy of the command output. In [10], a robotic arm was built according to voice and gesture commands given by an operator. If a speech was not recognized, gesture commands were used as an alternative. In addition, a five-layer multimode human-computer interaction framework was constructed in [11], comprehensively applying voice, text, motion perception, touch, and other human-computer interaction modes. The original data in the different modes were gradually mapped to various aspects during the execution of the system.

In contrast with these methods, the method proposed in this paper does not assume the dominant input modes or map the input modes to the output results. The proposed method

employs a decision-level approach (late fusion) that dynamically weights the information from a single-mode after a recognizer has interpreted the data. This strategy is typical in human–computer interaction systems because it can be easily extended and adapted.

### 3. Multimodal Interaction Integration System

This paper selects two of the most common human–computer interaction methods. The voice-interaction method involves using voice-recognition technology to understand the meanings of command words. The gesture interaction method uses a Leap Motion controller with infrared lighting and two grayscale cameras to capture 3D data of the hands. This method determines hand actions based on continuous hand posture frames and establishes interaction between the hands and the computer. In view of the differences among the interaction modes and interaction characteristics, a multimodal human–computer interaction integration framework is proposed, as shown in Figure 1. The original contents of the different channels are mapped to the system to produce a high-accuracy instruction output.



**Figure 1.** The multimodal integration framework for fusing voice and gesture data.

#### 3.1. Voice Recognition

Voice recognition is completed by an offline command word recognition-based software development kit (SDK) provided by the iFlytek open platform. The user speaks operation instructions (i.e., command words) into a voice input device, such as a microphone. The iFly software compares the received command words with the preset grammar, identifies the specific command information according to the comparison results, and transmits the results to a multimodal fusion terminal. These results are used in the later decision fusion step.

A standard command system should have three constituent elements—subject, predicate, and object—to meet the development needs of modern operating systems. To improve the accuracy of the voice recognition results, the grammar rules of the user voice inputs in the UAV task control stage (Equation (1)) are determined according to a large number of voice input experiments. The specific BNF syntax file is shown in Table 1.

$$\begin{aligned} \text{Voice Input}(VI) = & \text{Object description}(OD) + \\ & \text{Action description}(AD) + \\ & \text{Degree description}(DD) \end{aligned} \quad (1)$$

Each complete command includes the target object, the action to be performed, and the scale of the action. A good UAV mission command in a VR environment should consist of the UAV in the swarm, the current viewpoint position, and the scene. The UAV and the viewpoint position can move in parallel at different scales. In addition, the UAV can yaw left or right at different angles. The scene includes the global scene and the following scene and can be zoomed in or out at different scales.

**Table 1.** The grammar rules of the user’s voice inputs in UAV task control.

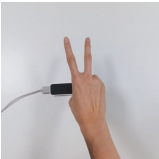
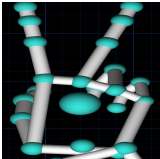
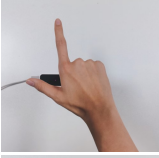
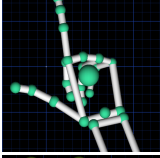
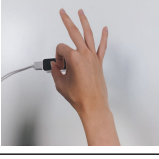
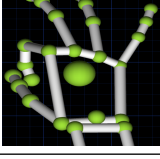
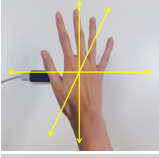
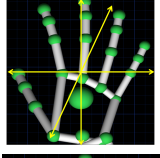
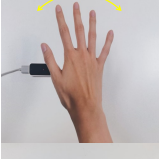
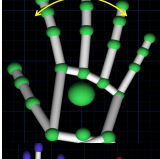
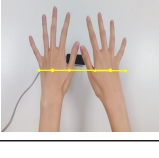
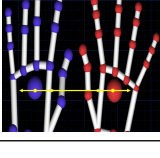
<b>BNF Syntax File for UAV Command Control</b>
!start <VI>;
<VI>::=[<OD>][<AD>][<DD>];
<OD>::=<UAVs>   <viewpoint>   <scene>;
<AD>::=<forward>   <back>   <left>...;
<DD>::=<large>   <middle>...;
<UAVs>::=[<id>]<UAV>;
<id>::=1   2   3...;
<UAV>::=UAV;
<viewpoint>::=current viewpoint position;
<scene>::=scene;
<forward>::=move forward;
<back>::=move back;
<left>::=move left;
<large>::=large scale;
<middle>::=medium scale;

### 3.2. Gesture Interaction

The Leap Motion controller consists of two high-definition cameras, three infrared LED lights, and optical sensors. The infrared LED light compensation mechanisms and dual high-definition cameras at different positions are used to obtain high-resolution infrared stereo images, which are then used to simulate human binocular stereo images to determine the associated gesture position [12]. A gesture recognition algorithm is proposed based on the physical Leap Motion controller model. The 3D information contained in the fingers, palms, elbows, and so on match the prebuilt physical models. The state of a motion, including the shape and the movement, can be determined according to the physical models. The gesture recognition process consists of four steps: acquiring the gesture data, preprocessing the feature vector data, extracting and optimizing the trajectory, and determining the trajectory recognition decision.

The Leap Motion sensors detect the fingers in each frame and use interface functions to capture the position and rotation angle of the hand, the speed of the fingers, etc., capturing hand movements with millimeter accuracy. Based on raw data that were previously collected with the Leap Motion controller application programming interface (API), we build features for recognizing gestures. To better integrate these results with the voice recognition results, the gesture design entities are divided into three types: the object, the action, and the degree. Furthermore, to minimize the impact of occlusion on the recognition confidence of the Leap Motion controller, clearer gestures are used as much as possible during the gesture design process. Shao [13], Fang [14], and Liu [15] provide a more general terminology for gesture control. Inspired by them, we designed a reasonable gesture database. Table 2 shows some of the gesture commands that are mapped to the robot functions. Among them, the velocity of the palm is calculated. If the movement value exceeds a predefined threshold, we assume that the hand is moving; otherwise, we classify the movement as a static gesture.

Table 2. Gesture recognition results

	Hand Posture	Recognition Results	Action description	Command Information
Object			Static gesture in which only the index and middle fingers are extended, while the other three fingers are bent.	UAVs
			Static gesture in which only the thumb and index fingers are extended, while the other three fingers are bent.	Scene
			Static gesture in which only the thumb and index fingers are extended, while the other three fingers are bent.	Viewpoint
Action			Five fingers extended and moving forward, backward, left, right, up, or down in parallel movements.	Moving forward, backward, left, right, up, or down
			All five fingers extended and rotating vertically.	UAV yaws to the left or right
			The five fingers of both hands are extended, and the distance between the hands gradually increases or decreases.	Zooming in or out of the scene.
Degree	–	–	The distance between the fingertips is small or large.	Ten or a thousand degrees
	–	–	The yaw rotation angle is small, medium, or large.	30 degrees, 45 degrees, or 60 degrees

### 3.3. Multimodal Fusion Interaction

A large number of studies have proven the excellent performance of various voice and gesture recognition methods; however, single modal recognition technology has some limitations. In this paper, the weighted fusion method is used to combine the recognition results of the voice and gesture modes; thus, the results complement and corroborate each other, improving the security and robustness of the command output.

Suppose that there are  $N$  predesigned commands  $C_i$  ( $i = 1, \dots, N$ ). The method for fusing the processing results of the voice and gesture inputs of these  $N$  commands can be formulated as follows. Because the voice recognition input and gesture recognition input have different processing periods, the longer voice recognition period is selected as the fusion baseline period, which is denoted as  $T_\alpha$  and defined explicitly in Equation (2).

$$T_\alpha = T_{\alpha\_end} - T_{\alpha\_start} \tag{2}$$

$T_{\alpha\_start}$  is the start time of voice recognition, and  $T_{\alpha\_end}$  is the generation time of the voice recognition result.

In a cycle  $T_\alpha$ ,  $pv_i$  represents the confidence degree of the voice recognition process as a command  $C_i$ . Assume that the number of gestures recognized as command  $C_i$  is  $m_i$  and that the confidence degree of each gesture recognition result is  $ph_{i,j}$  ( $j = 1 \dots m_i$ ). Then, the set of gesture recognition results can be defined as Equation (3).

$$\Gamma = \left\{ (C_1, ph_{1,1}), \dots, (C_1, ph_{1,m_1}), \dots, (C_N, ph_{N,1}), \dots, (C_N, ph_{N,m_N}) \right\} \quad (3)$$

In  $\Gamma$ , the ratio  $rh_i$  of the gestures recognized as command  $C_i$  is:

$$rh_i = \frac{m_i}{\sum_{j=1}^N m_j}. \quad (4)$$

The average confidence coefficient  $eph_i$  across all gestures recognized as command  $C_i$  is:

$$eph_i = \frac{\sum_{k=1}^{m_i} ph_{i,k}}{m_i}. \quad (5)$$

Then, the total confidence  $x_i$  of the gestures recognized as command  $C_i$  in the current cycle  $T_\alpha$  is defined as:

$$\begin{aligned} x_i &= rh_i \times eph_i \\ &= \frac{m_i}{\sum_{j=1}^N m_j} \times \frac{\sum_{k=1}^{m_i} ph_{i,k}}{m_i} = \frac{\sum_{k=1}^{m_i} ph_{i,k}}{\sum_{j=1}^N m_j}. \end{aligned} \quad (6)$$

Since each submodule has a different model, the forms of the resulting confidence results differ. In this paper, the total confidence of the gestures recognized as command  $C_i$  is used to investigate the similarity between the observed data and the internal model, as well as the number of recognition iterations. In contrast, the confidence of the voice recognition results involves estimating the posterior probability that the recognition result  $C_i$  is generated by the speaker. Notably, these two confidence values cannot be used directly as the input weights in the fusion system; thus, before the fusion operation, the two confidence values must be normalized.

$x_i$  is normalized to obtain  $\tilde{x}_i$  as follows:

$$\tilde{x}_i = \frac{x_i}{\sum x_i} = \frac{\sum_{k=1}^{m_i} ph_{i,k}}{\sum_{i=1}^N \sum_{k=1}^{m_i} ph_{i,k}}. \quad (7)$$

Then, the voice recognition confidence  $pv_i$  is normalized as  $\tilde{pv}_i$ :

$$\tilde{pv}_i = \frac{pv_i}{\sum pv_i}. \quad (8)$$

Therefore, the fusion probability  $\phi_i$  of the voice and gesture inputs of command  $C_i$  in cycle  $T_\alpha$  is calculated as follows.

$$\phi_i = \tilde{x}_i + \tilde{pv}_i \quad (9)$$

Among the object, action, and degree commands, the command  $C_i$  with the largest  $\phi_i$  value is selected. This value is adopted as the final command recognition result set ( $C_{Object}, C_{Action}, C_{Degree}$ ) in the current period  $T_\alpha$ .

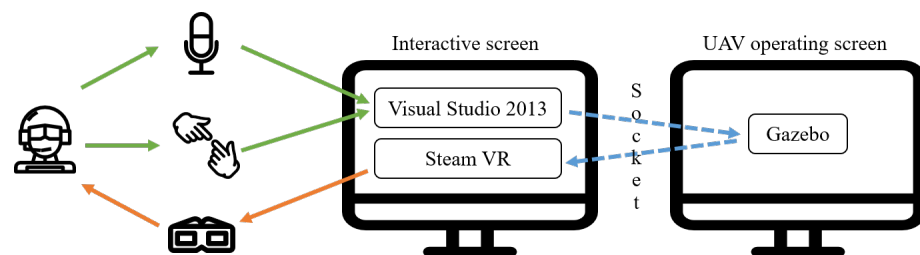
#### 4. Experiments

In this section, tests are conducted with the multimodal interaction integration system. The accuracy test evaluate the effectiveness of multimodality; the human factor ergonomics test restores real scenarios and evaluates the practical performance of the tested method.

The hardware environment includes the following components: a VR headset, a set of gesture interaction equipment, a set of voice-input equipment, and two personal computers,

one with Windows operating system, including Visual Studio 2013 and Steam VR software, for human–computer interaction command and control; the other with Ubuntu operating system, responsible for running Gazebo software in the robot operating system (ROS), for UAV simulation, which can generate UAV flight data based on users' input simulation.

As shown in Figure 2, the process of deploying the whole simulation experiment on the platform is as follows. First, the interactive devices capture the gesture and speech information. Second, the system fused them on the Windows system and the fusion result is converted into command information. Third, it was passed to the ROS in the Ubuntu system via socket programming with user datagram protocol (UDP) in real time. Fourth, the operating system returns the current position coordinates of the UAV and converts them to geodetic coordinates. Finally, Steam VR displays the UAV moving process loading on the VR device.



**Figure 2.** The process of deploying simulation.

#### 4.1. Accuracy Test

This article investigates voice interactions, gesture interactions, and multimodal interactions. Random combinations of the defined objects, actions, and degree commands are designed, yielding 34 sets of complete commands. The experiment uses five tests for each set of full commands; thus, a total of 170 sets of experiments are designed. Each speech command can be controlled in less than 5 seconds, and the gesture interactions and voice commands are output simultaneously.

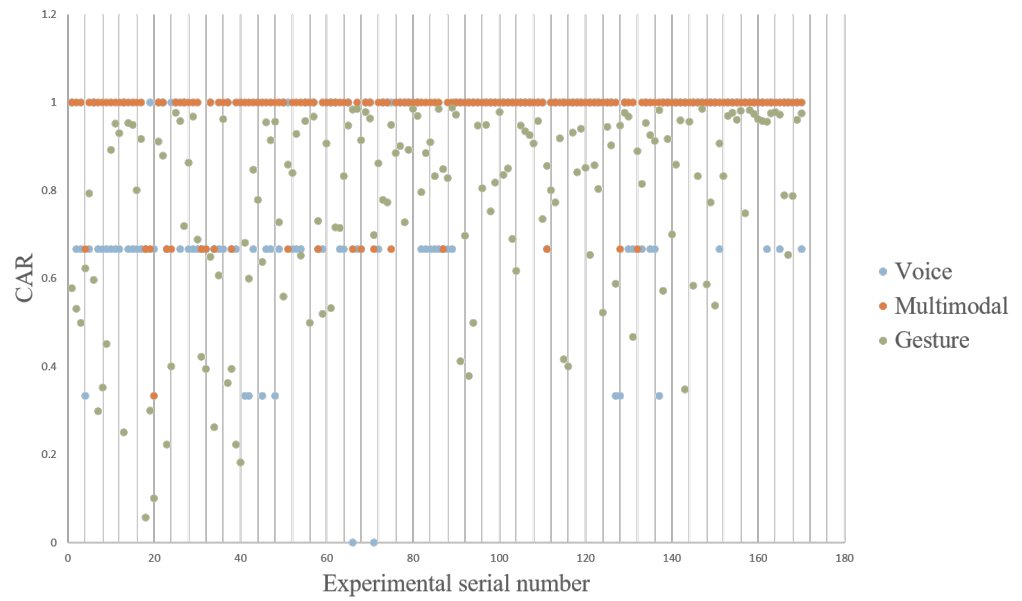
To compare the effect of the multimodal fusion method proposed in this paper with single-mode gesture recognition and single-mode voice recognition, we compared the accuracies of the command results obtained with the different interaction methods under the same input conditions. Referring to the industry-standard definition of the word error rate (*WER*) for measuring speech accuracy [16], we define the command result accuracy rate (*CAR*) as follows:

$$CAR = 1 - \frac{I + D + S}{N}. \quad (10)$$

To calculate the *CAR*, we first calculate the number of error commands identified during the recognition process and then divide this value by the number of recognition commands marked as artificial (*N*). The error identification commands can be classified into three categories:

- Insert (*I*): A command that is incorrectly added to the hypothetical script.
- Delete (*D*): A command that is not detected from the hypothetical script.
- Substitution (*S*): A command that replaces the hypothetical script.

The complete experimental results are shown in Figure 3. As the *CAR* value approaches 1, the interaction recognition result becomes more accurate. An error occurs in the experiment when a gesture or voice is missed or captured incorrectly. With the exception of a few cases, the multimodal fusion recognition-based method proposed in this article is more accurate than the single-modal recognition methods under the same input conditions. The accuracy of the multimodal fusion-based recognition method reaches as high as 100%. In particular, we analyze the rare cases in which the accuracy of the multimodal recognition method is less than that of the single-modal recognition methods. These errors occur mainly when the *pv;s* of the speech recognition commands are 0 and the wrong gestures are read; e.g., a horizontal hand movement is sometimes identified as a yaw command. Overall, the multimodal fusion of the voice and gesture data significantly improves the system performance.



**Figure 3.** The command result accuracies achieved by the different interaction methods under the same input conditions.

The average accuracy of each interaction method is defined as  $\widetilde{CAR}$ . This value is used to describe and compare the effects of the various interaction methods.

$$\widetilde{CAR} = \frac{\sum CAR}{EN} \times 100\% \tag{11}$$

where  $EN$  indicates the total number of experiments. In this experiment,  $EN$  is equal to 170. The experimental results are shown in Table 3.

**Table 3.** The average accuracies of several interaction methods.

Interaction Methods	Gesture	Voice	Multimodal
$\widetilde{CAR}$	76.9502%	83.8264%	95.8824%

After the voice and gesture data are fused, the average accuracy reaches 95.8824%. This result is noteworthy, as the accuracy is 12.056% better than that of the single modal speech recognition method and 18.9322% better than that of the single modal gesture recognition method. The above experimental results prove that the multimodal fusion of voice and gesture data significantly improves the command recognition accuracy of the proposed approach.

Considering the differences between experimental devices, we mainly consider the fault tolerance of the multimodal fusion method when conducting the comparison experiments; i.e., we hope to obtain highly accurate fusion results even when the accuracies of voice and gesture recognition are low. To facilitate a quantitative analysis, as shown in Equation (12), we define the fault tolerance index as  $CAR_{compare}$ .

$$CAR_{compare} = \frac{CAR_{multimodal}}{CAR_{gesture} + CAR_{voice}} \tag{12}$$

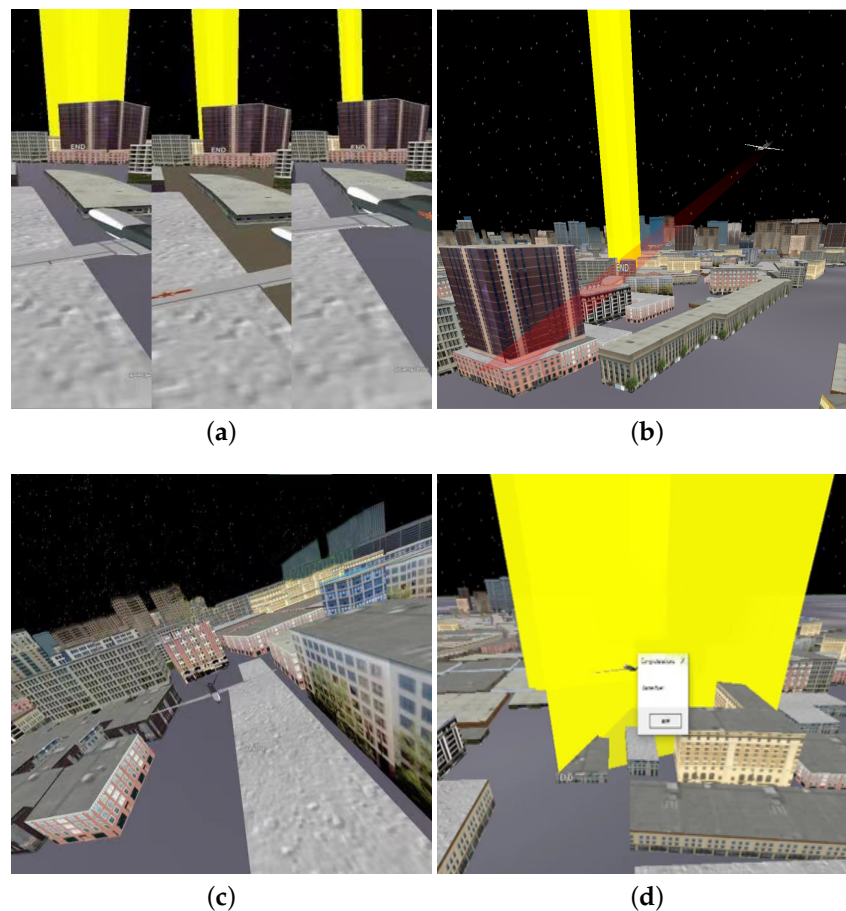
where the various parameters indicate the average gesture, voice, and multimodal accuracies. Notably, the larger the index, the higher the fault tolerance level of the system. In [9], the value of  $CAR_{compare}$  is 56.89%, while the value of our fusion method is 59.63%, which is 2.74% higher, so our method is valuable for improving the accuracy of the fusion results in the decision layer.



#### 4.2. Human Factor Ergonomics Test

To test the system's performance, we set up a simple urban environment exploration scenario with the UAV's initial position and different ranges of target positions in advance. The volunteers wore a pico neo headset, equipped with Leap Motion and a microphone, and output commands through gestures and voice interaction to control the UAV moving from the initial location to the target location in an immersive urban environment to explore the urban environment and obstacle avoidance. The measurement process allows volunteers to repeatedly maneuver the UAV for urban environment exploration walkthroughs and record their performance.

Ten non-professional volunteers of different ages and genders participated in the system evaluation. The experiment was divided into three phases. In the first phase, the volunteers' personal information related to the multimodal interaction integration system and their expectations were obtained. The volunteers also studied and became proficient in the system for 15–30 min. Next, in the second phase, the volunteers used the multimodal interaction integration system to perform six urban environment exploration tasks with set starting and target locations, including setting up three groups of experiments with different endpoint target ranges ( $20\text{ m} \times 20\text{ m}$ ,  $40\text{ m} \times 40\text{ m}$ ,  $80\text{ m} \times 80\text{ m}$ ), and conducted each group of experiments twice, to keep records and for the assessment of task performance; part of the experimental process is shown in Figure 4. In the third phase, a questionnaire evaluated the usability of the system, in which the volunteers score the questions on a seven-point scale, and the data were collected for data analysis and summarization.



**Figure 4.** Intermediate results of operating the UAV for urban exploration. (a) The difference of three target areas. (b) The trajectory of the UAV moving towards the target point. (c) The following viewpoint of the UAV. (d) The UAV approached the target point.

Figure 4a shows the difference in visual effects of three target areas, and the volunteer determined the location of the target point according to the yellow flashing alert. After that, the volunteer sent out UAV-related instructions through the multimodal interactive system, changed the speed of the UAV in each direction, and controlled the UAV to move toward the target point. The red part in Figure 4b is the UAV trails. While operating, the volunteer could switch the following viewpoint to observe the specific location of the drone and the surrounding environment information, as seen in Figure 4c. If there was a building, the volunteers could adjust the direction of travel in time, and eventually, the UAV approached the target point along a safe path. After reaching the target area, a pop-up window of the end of the game appeared, as seen in Figure 4d.

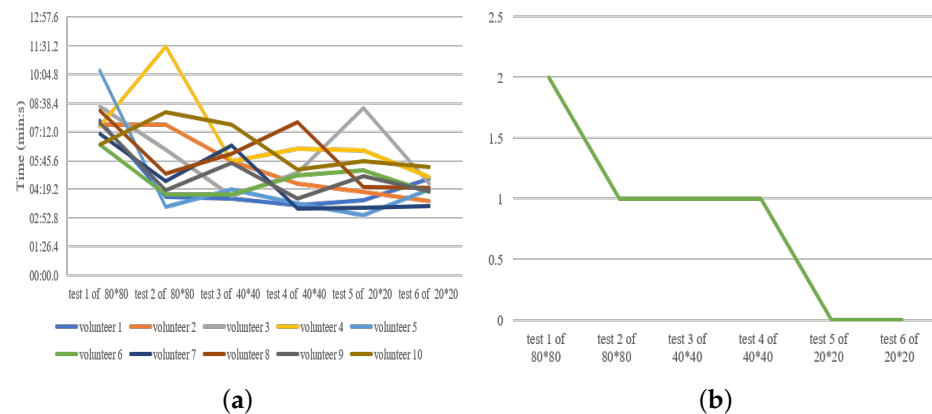
In order to verify whether the system could assist the operator to complete the task target and improve the manipulation efficiency, as well as whether the command output was sensitive, we counted the number of commands per volunteer and the number of incorrectly identified commands during the volunteer's manipulation of the UAV movement. Additionally, some of the experimental results are shown in Table 4. Different volunteers had various ways to reach the destination point and the frequency of command output due to different habits, but false recognition is minimal in relation to the total number of commands, which means that there is overall high recognition accuracy, which verifies the effectiveness of the system.

**Table 4.** The fusion output of different volunteers.

	Up		Down		Forward		Backward		Left		Right	
	All	Fault	All	Fault	All	Fault	All	Fault	All	Fault	All	Fault
volunteer 1	15	0	11	0	49	0	0	0	2	0	23	0
volunteer 2	13	1	10	0	31	0	0	0	0	0	13	0
volunteer 3	15	0	16	0	28	3	0	0	10	0	24	0
volunteer 4	12	1	13	0	35	2	0	0	1	0	17	0
volunteer 5	21	0	15	0	40	0	0	0	3	1	20	0
volunteer 6	13	1	15	0	31	2	0	0	1	0	14	0
volunteer 7	17	1	12	0	21	0	0	0	5	0	9	1
volunteer 8	17	0	16	0	30	1	3	0	6	0	23	0
volunteer 9	12	0	10	0	36	2	0	0	2	0	14	1
volunteer 10	12	1	12	0	24	0	0	0	4	0	15	0
the overall	147	5	130	0	325	10	3	0	34	1	172	2

Meanwhile, Figure 5a shows the time consumed by each volunteer to reach the target point each time and Figure 5b shows the number of collisions with the building. From the experimental results, despite the gradual reduction of the target range, the time spent on the operation still gradually decreases as the number of user drills increases, and the collision with the building occurs very few times and all occur in the first few experiments. Very few abnormal cases, such as the third volunteer in the fifth experiment, took considerably more time, which indicates that the system is designed to be simple and easy to interact with, and has low learning costs, which can effectively improve the user's familiarity with the urban environment and enhance the efficiency of UAV control.

In this paper, we designed a questionnaire (Table 5) based on VRUSE [17] about the system's overall effectiveness, covering seven aspects: completion, functionality, interactivity, consistency, learnability, usability, and comfort. Then we statistically analyzed the results of the questionnaire. The mode and average values are above 6, the volunteers mostly agree with the descriptions of the questions in Table 5, and the variance is less than 0.5, which reflects the consistency of the volunteers' evaluations. The results indicate that the volunteers found the multimodal interaction integration system is usable.



**Figure 5.** Task performance. (a) The time consumed by each volunteer to reach the target point in the six tests. (b) The total number of hit buildings in each test for all volunteers.

**Table 5.** The questionnaire and scores of availability measures.

Category	Detail	Mode	Average	Variance	
1	Completion	The system supports volunteers in performing the complete task of manning a UAV for urban environment exploration.	7	7	0
2	Functionality	The system’s method of completing tasks, frequency of operation, and duration are reasonable.	7	6.9	0.09
3	Interactivity	The system allows volunteers to interact naturally and flexibly with the scene and the UAV in real-time, and where attention is focused more on the task than on the interactive interface and physical devices.	7	6.8	0.16
4	Consistency	The system responds interactively in the way the volunteer expects, and the feedback is clear and easy to understand.	7	6.7	0.41
5	Learnability	The system is intuitive and easy to learn.	7	6.8	0.16
6	Usability	The system is intuitive and easy to use.	7	6.8	0.36
7	Comfort	The use of helmets, gestures, and voice control in the system did not make volunteers feel fatigued, nauseous, vertigo, or other discomforts.	7	6.7	0.21

However, they also found some shortcomings of the system. The recognition system is not very friendly to people with short and thick fingers and accents, the command recognition speed is slow, the voice-recognition lexicon is small, and the gesture recognition may be interfered with by irrelevant postures. Therefore, in the future, we must integrate more modules in our system, expand the speech recognition library, and exclude the interference items in gestures. Furthermore, due to there being more models in the scene, the frame rate is reduced, which may easily cause visual discomfort to users. Therefore, the improvement of visualization is also our next research direction.

### 5. Conclusions

This paper proposes a multimodal interaction integration system that realizes accurate UAV control command outputs in virtual environments by fusing gesture and voice data. In the multimodal data fusion process, we propose a new interaction method and adjust the interaction strategy according to the confidence of the single-modal recognition result; that is, we dynamically modify the fusion weight, providing intelligent interactive service support for command recognition. We also evaluate and compare the single-modal

and multimodal fusion-based interaction recognition results obtained with this system. The accuracy test results show that the multimodal interaction integration system that combines gesture and speech data is more accurate than the single interaction modalities. Then the human factor ergonomics test results indicate the system's overall effectiveness and usability. The next step is to integrate more natural interaction modules and improve visualization.

**Author Contributions:** Conceptualization, X.X. and Q.T.; methodology, X.X. and Q.T.; software, X.X. and Q.T.; validation, Q.T.; formal analysis, Q.T.; investigation, Q.T.; writing—original draft preparation, X.X. and Q.T.; writing—review and editing, H.Z., D.T. and J.L.; supervision, H.Z., D.T. and J.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank the Swarm Observation and Regulation Laboratory, National University of Defense Technology, Hunan Province, China, for the resources provided by them.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Alexander, G.H.; Paul, M. Gestures with speech for graphic manipulation. *Int. J. Man-Mach. Stud.* **1993**, *38*, 231–249.
- Kelly, S.D.; Özyürek, A.; Maris, E. Two Sides of the Same Coin: Speech and Gesture Mutually Interact to Enhance Comprehension. *Psychol. Sci.* **2010**, *21*, 260–267. [[CrossRef](#)] [[PubMed](#)]
- Lee, M.; Billinghamurst, M.; Baek, W.; Green, R.; Woo, W. A usability study of multimodal input in an augmented reality environment. *Virtual Real.* **2013**, *17*, 293–305. [[CrossRef](#)]
- Geraint, J.; Nadia, B.; Roman, B.; Simon, J. Towards a situated, multimodal interface for multiple UAV control. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Anchorage, Alaska, 3–8 May 2010; pp. 1739–1744.
- Kensho, M.; Ryo, K.; Koichi, H. Entertainment Multi-rotor Robot that realises Direct and Multimodal Interaction. In Proceedings of the 28th International BCS Human Computer Interaction Conference (HCI), Southport, UK, 9–12 September 2014; pp. 218–221.
- Jane, L.E.; Ilene, L.E.; James, A.L.; Jessica, R.C. Drone & Wo: Cultural Influences on Human-Drone Interaction Techniques. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, Colorado, USA, 6–11 May 2017; pp. 6794–6799.
- Clifford, R.; McKenzie, T.; Lukosch, S.; Lindeman, W.R.; Hoermann, S. The Effects of Multi-sensory Aerial Firefighting Training in Virtual Reality on Situational Awareness, Workload, and Presence. In Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Atlanta, GA, USA, 22–26 March 2020; pp. 93–100.
- Burger, B.; Ferrané, I.; Lerasle, F.; Infantes, G. Two-handed gesture recognition and fusion with speech to command a robot. *Auton Robot.* **2012**, *32*, 129–147. [[CrossRef](#)]
- Mohd, T.K.; Carvalho, J.; Javaid, A.Y. Multi-modal data fusion of Voice and EMG data for Robotic Control. In Proceedings of the 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), Columbia University, NY, USA, 19–21 October 2017; pp. 329–333.
- Kandalaf, N.; Kalidindi, P.S.; Narra, S.; Saha, H.N. Robotic arm using voice and Gesture recognition. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), UBC, Vancouver, BC, Canada, 1–3 November 2018; pp. 1060–1064.
- Chen, Y.; Zhao, H.; Chen, J. The Integration Method of Multimodal Human-Computer Interaction Framework. In Proceedings of the 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 11–12 September 2016; pp. 545–550.
- Huang, J.; Jing, H. Research on gesture recognition in virtual interaction based on Leap Motion. *Appl. Res. Comput.* **2017**, *4*, 1231–1234.
- Shao, L. *Hand Movement and Gesture Recognition Using Leap Motion Controller*; International Journal of Science and Research: Raipur, Chhattisgarh, India, 2016.
- Fang, X.; Wang, Z.; Gao, J. Design of Gesture Interaction System Based on Leap Motion. *Mach. Des. Res.* **2020**, *36*, 128–132.
- Liu, Y.; Wang, S.; Xu, G.; Lan, W.; He, W. Research on 3D Gesture Interaction System Based on Leap Motion. *J. Graph.* **2019**, *40*, 556–564.

- 
16. Negrão, M.; Domingues, P. SpeechToText: An open-source software for automatic detection and transcription of voice recordings in digital forensics. *Forensic. Sci. Int. Digit. Investig.* **2021**, *38*, 301223. [[CrossRef](#)]
  17. Kalawsky, R.S. VRUSE—A Computerized Diagnostic Tool: For Usability Evaluation of Virtual Synthetic Environments Systems. *Appl. Ergonomics* **1999**, *30*, 11–25. [[CrossRef](#)]