IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 28, 2020, accepted September 15, 2020, date of publication September 18, 2020, date of current version October 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025229

A Design of Smart Beaker Structure and Interaction Paradigm Based on Multimodal Fusion Understanding

DI DONG, ZHIQUAN FENG, (Member, IEEE), JIE YUAN, XIN MENG, JUNHONG MENG, AND DAN KONG

Department of Information Science and Engineering, University of Jinan, Jinan 250022, China
Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China

Corresponding author: Zhiquan Feng (ise_fengzq@ujn.edu.cn)

This work was supported by the National Key R&D Program of China (No. 2018YFB1004901), and the Independent Innovation Team Project of Jinan City (No. 2019GXRC013).

ABSTRACT Virtual chemistry experiment is an important teaching tool in middle schools. It can not only help students understand the experimental principles, but also help them memorize the experimental procedures. However, there are still some problems in virtual chemistry experiments. First, in existing research, user intentions are often misunderstood and cannot be accurately understood. Second, the existing research fail to identify the user's wrong actions, which reduce the accuracy of the experiment. Third, the user sense of operation and realism is not strong during the experiment, which reduces the user's experience. Finally, the lack of navigation guidance for experimental operations increases user learning time. In order to solve these problems. This paper proposes a scheme for establishing an intelligent navigational chemical laboratory based on multimodal fusion. First of all, we design a new smart beaker structure with perceptual ability, which can be used to complete most chemical experiments and give users a real sense of experience. Besides, we propose a multimodal fusion understanding algorithm, which reduces the misidentification of the experiment and better understands the real intention of the user. Finally, intelligent navigation and wrong behavior recognition functions are added to the experimental equipment, which improves the efficiency of human-computer interaction. The results show that Compared with the existing virtual laboratory or system, the chemical laboratory scheme proposed in this paper through multimodal fusion understanding algorithm greatly reduces the user's memory load and improves the success rate of the experiment. Moreover, through the combination of virtual and real, the virtual chemistry experiment not only improves the authenticity of the operation, but also stimulates the students' interest in learning, which is well received by users.

INDEX TERMS Human–computer interaction, multimodal fusion, intention understanding, smart beaker, navigation interaction paradigm.

I. INTRODUCTION

Many of the experiments in middle school chemistry textbooks are destructive, costly, and relatively dangerous. As a result, many teachers have simply removed these experiments, and students memorize only the chemical equations and reactions. Therefore, students are prone to a variety of problems such as weak memory of knowledge points, confusion, and lack of practical hands-on ability. With the development of modern technology and the application of numerical simulation technology, the virtual simulation experiment system solve this problem. Many chemistry experiments can be completed in a virtual simulation system. However, most of the current virtual simulation experiment systems use a mouse and keyboard as input, and the output is displayed. Although it can be used to solve many chemical experiments that are difficult to actually operate, it greatly limits the user's operation and cannot give the user a real sense of feedback. And most systems lack the ability to understand users' intentions and the ability to navigate user behaviors, which increases the user's operating load and reduces the efficiency of the experiment.

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Tucci.

Therefore, this paper aims to perform chemistry experiments that are difficult to operate, involve high risk factors or expensive reagents in a virtual chemistry experiment environment. Then, we want to give students more realistic scene perception feedback and Enable students to complete chemical experiments without burden and efficiently. First, we take 'the combination of theoretical teaching and practical teaching, the combination of physical experiment and virtual simulation [1] as the design goal, and the design orientation of 'multimodal fusion and intention understanding' to construct a simulated chemistry experiment based on a fusion of speech recognition and behavior sensors, so that the user can be out of the control of the mouse and keyboard, and gain the real sense of operation. In addition, we use the combination of sensor and voice in multimodal fusion and intention understanding. So, users can use it as a real experiment, and greatly reduce the user's memory load. Finally, we use the navigation interaction paradigm, which greatly improves the intelligence of the system so that users do not need to deliberately remember the steps, and the operation is more natural and convenient.

The main contributions of this paper are as follows:

1. We use 3D printing to design a set of smart beaker models with perceptual ability. Users can complete most of the middle school students' chemistry experiments by operating this smart beaker.

2. We propose a multimodal fusion understanding algorithm, through which the user's behavior can be perceived and intention understanding can be realized, and misidentification can be reduced.

3. We propose a navigational experimental system interaction paradigm, which shortens the time for users to understand the experimental process, and proposes a feasible solution to solve the problem of imbalance of teachers in remote areas of central and western China and lack of teacher guidance.

II. RELATED WORK

A. VIRTUAL TEACHING

With the development of computer technology, virtual teaching has gradually entered the teaching environment of schools. And scholars at home and abroad have also begun to conduct research to promote the development of virtual teaching.

Zhang *et al.* [1] studied the influence of a network robot system (NRS) on virtual boundary teaching and proposed a method based on an NRS and a laser pointer as interactive devices. This interactive approach can integrate the robot into a smart environment and support the teaching process in terms of interaction and feedback. The results showed that this system can achieve the equal success and accuracy compared to methods that do not support NRS and has shorter teaching time. Peixoto *et al.* [3] proposed a method of using VR technology for virtual foreign language teaching and evaluated the benefits of this method as a tool to help students with listening tasks. The results showed that this method is not only attractive but also can help motivate students. Yu [4] proposed a two-channel classroom feedback system integrated with a back-end, e-learning system. The system interacted through a two-channel mechanism, gained a teaching response by identifying verbal keywords, and achieved a social response by analyzing the social signals provided by the student's head movements. The results showed that the dual-channel mechanism not only can provide the basic functions of a classroom feedback system but also greatly enhance the interaction between teaching practice and learning activities. So, students gained a better learning experience and more satisfaction. Wang *et al.* [5] developed a virtual chime-bell experimental system based on multimodal fusion for the increasing requirements of college classes, science and technology museums, etc. The results indicated that the multimodal fusion method can achieve a better performance with a virtual chime-bells experimental system than with other systems. Park *et al.* [6] researched the emotional response of preservice teachers with the virtual scene-based teacher training system, and created three types (no interaction, unexpected interactions and expected interactions) of interaction scenarios through SimTEACHER. Studies showed that participants have higher positive and neutral emotions, higher emotional engagement, and a higher sense of happiness when performing unexpected interactions than expected interactions. Adam *et al.* [7] conducted a study on how to apply virtual teaching in higher education institutions (HEIs) in developing countries. The authors believed that virtual teaching should be adapted to the background of developing countries in HEIs by modifying open source technologies and rules. Moreover, the study took a physics teaching environment as an example to study how to solve contradictions in the tools, implementers and rules in virtual teaching. Liu *et al.* [8] studied the potential, application and challenges of VR technology in the field of education based on the VRSC model. Yang *et al.* [9] analyzed the practical framework, application progress, development trends and realistic challenges of smart teaching and constructed teaching big data from four aspects: the introduction, construction, application and data driving educational big data projects. Al-Khalifa *et al.* [10] proposed a virtual chemistry laboratory application based on Leap motion. However, this system requires memorizing some complex gestures, which makes it difficult to use well. The authors of [11], [12] used intelligent equipment to conduct intelligent experiments. However, the existing virtual laboratories lack personalized functions. Ghergulescu *et al.* [13] introduced an interactive scientific virtual laboratory architecture, enabling students to create experiments and to practice at their own pace. With the help of an artificial intelligence paradigm, Munawar *et al.* [14] proposed an intelligent virtual laboratory using a pedagogical agent-based cognitive architecture.

Through the analysis of the above literature, it can be seen that applying virtual teaching to the teaching of chemical

experiments can enable students to experience experiments that cannot be carried out due to safety and other reasons, so as to make it easier for students to understand the mechanism of chemical reactions and improve their learning interest and efficiency.

B. MULTIMODAL FUSION AND INTENTION UNDERSTANDING

As the subjects of virtual experiment teaching, teachers and students have their own intention. Only by correctly understanding the intention can the experiment be carried out smoothly and efficiently. In the field of human-computer interaction, multimodal fusion has become an important research content of scholars at home and abroad in order to make computers understand users' intentions.

Wu *et al.* [10]–[18] thought that there are three fusion strategies: prefusion, post fusion and hybrid fusion. Atrey *et al.* [19] noted that fusion methods can be divided into three categories: rule-based, classification-based, and estimation-based methods. Qi *et al.* [20] used a priori information as a reference for guiding multimodal data fusion and proposed a fusion reference model, which can accurately identify covariant multimodal imaging modes closely related to the reference information. Choi and Lee [21] proposed a new deep learning-based multimodal information architecture, EmbraceNet, which improved the architecture of neural network feature extraction and classification. It can not only integrate multimodal related data but also effectively address missing data. Guo [22] studied multimodal learning and sensor fusion from the perspective of latent variables and proposed the first algorithm to solve online multimodal decision problems so that different types of sensors can make decisions together. Mi *et al.* [23] proposed a deep neural network structure based on a convolutional neural network (CNN) for human-machine intentional semantic understanding. Moreover, a multimodal interaction framework based on object availability was proposed to understand and execute semantics. Experiments showed that the framework is feasible and practical. Ehatisham-Ul-Haq *et al.* [24] proposed a new feature-based multimodal fusion method for the fusion of data from multiple sensors, including RGB cameras, depth sensors, and wearable inertial sensors. The results showed that the feature layer fusion of RGB and inertial sensors provided the best overall performance for the system, resulting in better recognition results than previous methods. Chung *et al*. [25] studied sensor and data acquisition by human activity recognition (HAR) systems and the multimodal fusion of activities of daily life (ADL) identified by different sensors. Furthermore, by analyzing the accuracy of each sensor's identification of different types of activities, the weights of each multimodal sensor were defined to more accurately reflect the characteristics of individual activities. Yang *et al.* [26] proposed a comprehensive probability decision framework based on distance-based inference (DBI) and knowledge-based inference (KBI), which was used by robots to infer the role of human beings in specific tasks. Besides, an information entropy-based weighted fusion scheme was used to estimate the behavior and location of a person in a target role. They found that this framework can make reasonable estimates of various situations. Jiang *et al.* [27] used subspace-based low-level feature fusion of face poses and speech to recognize specific speakers in human-computer interactions. Peruffo Minotto *et al.* [28] proposed a method for extracting speech features by an online multimodal algorithm by using a color camera and a depth sensor as input streams. Liu *et al.* [29] used three modes (facial expression, voice and gestures) as input for a robot. At the feature level, all the features were converted into feature vectors for emotion recognition. At the decision-making level, the naive Bayes classifier for the three different modes was used to obtain the probability of the sentiment analysis result for a single mode. Choosing the most probable result as the final result, the robot responded to different emotions expressed by the human. Ghayoumi *et al.* [30] proposed a robust multimodal interaction framework that can realize the interaction between rescuers and flying robots by integrating voice and hand and arm posture data based on post fusion integration in the decision-making layer.

In short, using the interactive method of multimodal fusion can use one modality to make up for the problems about ''incomprehension'', ''incompleteness'' or ''misunderstanding'' of the intention generated by another modality. It can eliminate the ambiguity caused by only relying on one modal input information to understand the intention, so that the computer can understand the user's intention more accurately, and the process of human-computer interaction is more efficient and natural.

To sum up, in the existing research on virtual teaching, most of the experiments are conducted mainly through virtual simulation, and only through the simulation of the experimental process, whilch makes users lack the sense of reality and experience. At the same time, some studies use a single interaction method, or lack of understanding and feedback of user intentions, or lack of guidance and prompts for the experimental process, which will affect the user's sense of operation and the fluency of the experiment. Therefore, this paper adopts the form of combination of physical operation and virtual presentation. We design a set of intelligent beaker model and make it perceive user behavior through light sensitivity and sound induction. Android software package (APK) that controls experimental variables is designed according to the Android operating system. Second, the multi-peak fusion algorithm based on the fusion of the decision-making layer is used as the core algorithm of the experimental equipment to realize the fusion of speech recognition and behavioral perception information. Then, based on the model of intent understanding, the correct experimental interaction is ensured through the navigation interaction paradigm, and finally an intelligent chemistry experiment system is formed.

III. STRUCTURE DESIGN AND USER PERCEPTION METHOD OF SMART BEAKER

A. BEAKER DISPLAY DESIGN

We design a new structure and sensing mechanism of the smart beaker to improve the cognitive ability. The smart beaker set up a smart screen on a container model, and manually set the experimental product (solid/liquid) parameters put into the container on the touch screen display. Fig. 1 shows the display device of the smart beaker. Users can set up the experiment, including the weight, volume, temperature, concentration, and other required experimental conditions through the touch screen on the display. A photosensitive is placed at the inlet edge of the container model. Each photosensitive is numbered sequentially from 0 (i.e., 0, 1, 2,...) according to the position of the photosensitive at the inlet edge of the container model. When the user transfers a liquid from one experimental container to another, one of the experimental containers is pressed against the inlet edge of the other experimental container; when the user places a solid experimental product model into the experimental container using a tweezers model, dicing is required by pressing on the inlet edge of the experimental container [31]. At this point, the operation state is determined by identifying the number of occluded photosensitive sensors. An acoustic sensor is placed at the bottom of the container model, and the sound sensor is connected to the electronic chip.

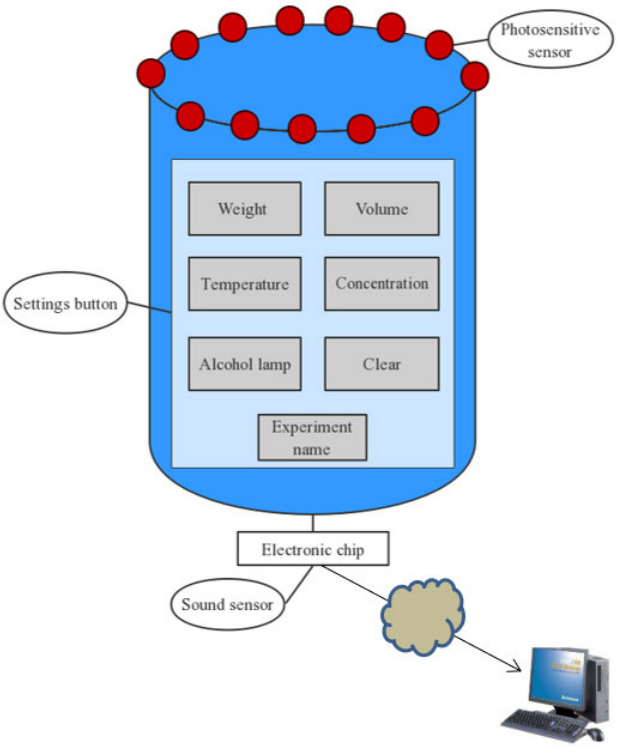


FIGURE 1. Calculation and display device.

The touch screen of this display is an Android APK that is run in the Android environment. First, the mobile terminal establishes a connection with the web server through the application programming interface (API). Next, the relevant data is encapsulated and sent to the web server in strict accordance with the Hypertext Transfer Protocol (HTTP) format. Then, the web server splits the information and passes it to the database [32]. The user sets the properties of the substance inside the beaker through the interactive interface of the Android terminal, including the weight of the substance and the name of the substance. And denote by *sw* and *sn* respectively. Finally, after *sw* and *sn* are stored in the database, the data are transmitted by reading the database system information through the computer system. The data transmission is shown in Fig. 2.

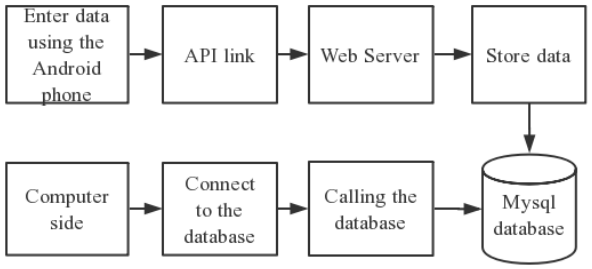


FIGURE 2. The data transmission of the touch screen display.

B. USER BEHAVIOR PERCEPTION METHOD

1) EXPERMENTAL SUBSTANCE PERCEPTION

In our experiment, we judge the operation of placing the experimental product based on the number of blocked photosensitive sensors. If the electronic chip detects a photosensitive signal, it calculates the number of photosensitive M ($M \leq N$) that are activated (i.e., the photosensitive resistance is transmitted to the electronic chip) and the sensor number i and j. $N+1$ is the total number of pressure-sensitive sensors on the container model [33]. The number of occluded photosensitive d is calculated by (1):

$$d = \underset{0 \leq i \leq j \leq M}{Max} (\|i - j\| \mod N + 1) \tag{1}$$

Taking into account the different properties of the substance added to the beaker in the chemical experiment (such as: solid substance, liquid substance). According to the characteristics of the operation behavior when adding these substances (such as: using tweezers to add solid substances, using container to pour liquid substances), we use the number of shielded photosensitive around the mouth of the beaker to determine the properties of the substance added. The relationship between the user's behavior of adding substances and the corresponding number of blocked photosensitive is shown in TABLE 1.

2) STIRRING BEHAVIOR PERCEPTION

When the user shakes the container model or stirs the liquid model with a glass rod, the sound produced by the contact of the glass rod or the magnetic stirrer on the inner wall of the container model is sensed by the sound sensor. We definition

TABLE 1. Experimental substance perception.

User behavior	Evaluation conditions (d)	Label (Ph)
Adding solid substances	$d=1$	$Ph1$
	$d=2$	$Ph2$
Adding liquid substance	$d=3$	$Ph3$
	$d=4$	$Ph4$

Note: Get the label according to d $Ph \in \{Ph1, Ph2, Ph3, Ph4\}$, It is used to judge the attributes of substances added by users.

the duration t_1 of the sound and the maximum amplitude f_1 of the audio during the duration. Then, we obtain the value of the empirical parameters τ_1 and κ_1. We set the thresholds v_1 and v_2 of the two speeds in advance to calculate the shaking speed v:

$$v = \beta f_1 \tag{2}$$

Here, the specific situation of the user's shaking or stirring behavior is determined according to the speed v calculated by the sound sensor signal. It is shown TABLE 2.

TABLE 2. Stirring behavior perception.

User behavior	Evaluation conditions (t_1, f_1, v)	Label (s)
Stirring	$t_1 > \tau_1$ and $f_1 > \kappa_1$	S
Stirring slow speed	$v \le v_1$	$S1$
Stirring fast speed	$v \ge v_2$	$S2$

Note: According to (t_1, f_1, v) get the label $s \in S = \{S1, S2\}$, It is used to judge the specific situation of the user's shaking or stirring behavior.

IV. NAVIGATIONAL INTERACTION PARADIGM BASED ON MULTIMODAL FUSION UNDERSTANDING

A. OVERALL FRAMEWORK

In this paper, an intelligent navigation interactive method based on intention understanding of multimodal fusion is designed for chemical experiments. This method uses a combination of physical interaction and virtual presentation (virtual and real fusion). Users observe the reaction phenomenon of the virtual scene while operating the smart beaker. By fusing sensor information and voice information to improve the understanding of user intentions, visual and auditory navigation prompts are added to solve the problems of perception and interaction in virtual teaching experiments. The overall framework is shown in Figure 3.

Navigational interaction paradigm based on multimodal fusion understanding consists of a five-layer structure. In the input layer, the user's voice or the information generated by several operations on the smart beaker will be recorded. In the perception layer, the information of the input layer is sensed through ordinary microphones, photosensitive, pressure sensors and sound sensors to obtain deeper feature information. Then, the recognition layer compares the feature information obtained by the perception layer with the keyword speech database and the operation behavior database to obtain the recognized speech intention information set and behavior intention information set. The two information sets will be used as the input of the fusion layer to understand the real intention of the user. Here, we propose a multimodal fusion understanding (MMFU) algorithm that can effectively and accurately obtain users' intention. Next, we propose a navigational interaction strategy. It is based on the purpose of solving chemical experiments in this paper and enables the system to perform visual presentation and auditory prompts of different situations according to the obtained intention. Finally, the application layer transmits the visual and auditory navigation that occurred to the user, and waits for the user's next behavior.

B. MULTIMODAL INFORMATION DATABASE

According to the function of the smart beaker and the characteristics of the input multimodal information, we establish the database of user behaviors for the system to improve the cognitive efficiency of the system. This database includes the behavioral intentions that users may have in the experiment. The database includes an operational behavior database and a keywords speech database.

In the operational behavior database, we define the information of photosensitive sensor, pressure sensor and sound sensor contained in the sensing device. As shown in TABLE 3.

TABLE 3. Operational behavior database.

Database name	Label	Set
Pho_1	1	$Pho\_1 \in \{Pho\_1, Pho\_2, Pho\_3, Pho\_4\}$
Pho_2	1	$Pho\_2 \in \{Pho\_1, Pho\_2, Pho\_3, Pho\_4\}$
Pho_3	2	$Pho\_3 \in \{Pho\_1, Pho\_2, Pho\_3, Pho\_4\}$
Pho_4	2	$Pho\_4 \in \{Pho\_1, Pho\_2, Pho\_3, Pho\_4\}$
Pes_1	3	$Pes\_1 = \{Pes\_1\}$
Sou_1	4	$Sou\_1 \in \{Sou\_1, Sou\_2\}$
Sou_2	5	$Sou\_2 \in \{Sou\_1, Sou\_2\}$

Note: in TABEL 3, $\{Pho\_1, Pho\_2, Pho\_3, Pho\_4\}$ are the photosensitive information set, which respectively indicate the data information when the number of photosensitive s blocked is 1, 2, 3, 4. They have a one-to-one correspondence with $\{Ph1, Ph2, Ph3, Ph4\}$ in TABLE 1. Pes_1 is the pressure sensor information set, which represents the force information when the sensor is pressed. $\{Sou\_1, Sou\_2\}$ are the sound sensor information set, which respectively represent the speed information when the sound occurs. They have a one-to-one correspondence with $\{S1, S2\}$ in TABLE 1. The value of Lable in the second column is represented by $labelb$.

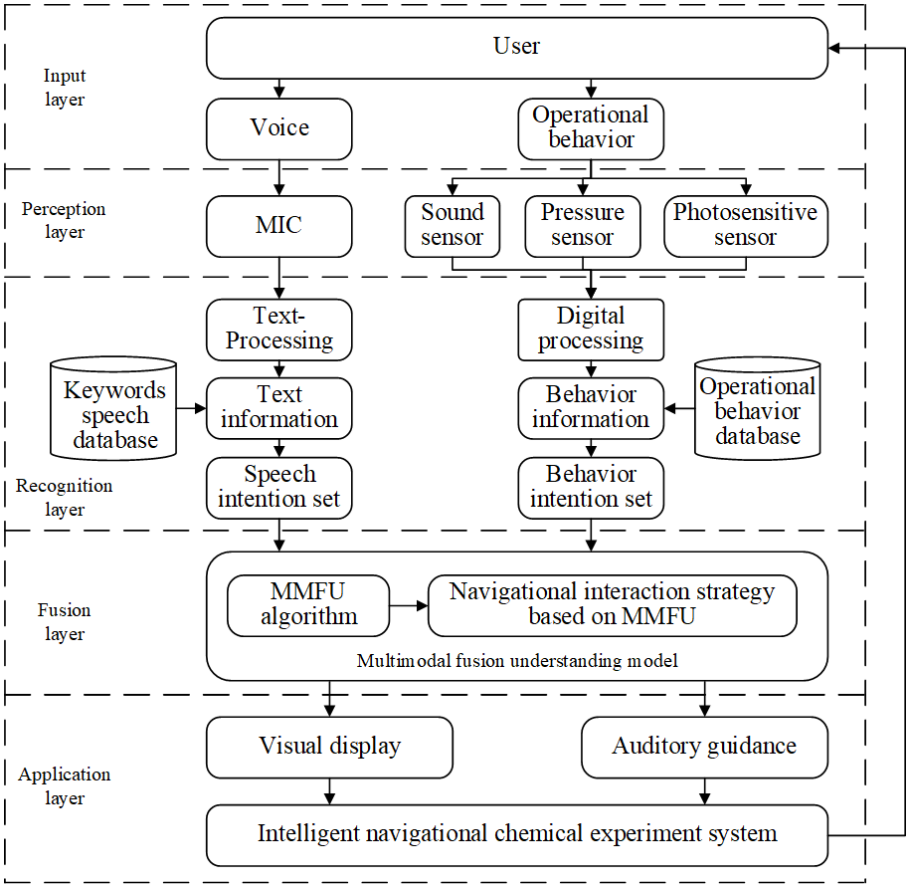


FIGURE 3. Overall framework.

In the keywords speech database, several keywords are assigned to each data set to match the text information input by the user. The definition of the specific data set is shown in TABLE 4.

TABLE 4. Keywords speech database.

Database name	Label	Set
Sp_1	1	$Sp\_1=\{Sp\_1_1, Sp\_1_2, Sp\_1_3, ..., Sp\_1_n\}$
Sp_2	2	$Sp\_2=\{Sp\_2_1, Sp\_2_2, Sp\_2_3, ..., Sp\_2_n\}$
Sp_3	3	$Sp\_3=\{Sp\_3_1, Sp\_3_2, Sp\_3_3, ..., Sp\_3_n\}$
Sp_4	4	$Sp\_4=\{Sp\_4_1, Sp\_4_2, Sp\_4_3, ..., Sp\_4_n\}$
Sp_5	5	$Sp\_5=\{Sp\_5_1, Sp\_5_2, Sp\_5_3, ..., Sp\_5_n\}$
Sp_6	6	$Sp\_6=\{Sp\_6_1, Sp\_6_2, Sp\_6_3, ..., Sp\_6_n\}$
Sp_7	7	$Sp\_7=\{Sp\_7_1, Sp\_7_2, Sp\_7_3, ..., Sp\_7_n\}$

Note: in TABEL 4, Sp_1, Sp_2, Sp_3, Sp_4, Sp_5, Sp_6, Sp_7 indicates the voice information set containing the keywords of "dumping", "removing solids", "dropping solution", "slow stirring or shaking", "fast stirring or shaking", "confirm", and "complete". The value of Lable in the second column is represented by *labels* .

C. MULTIMODAL INFORMATION PERCEPTION AND RECOGNITION

In the perception of multimodal information, sensing devices such as photosensitive sensors, pressure sensors, and sound sensors are used to perceive the user's operation behavior on the beaker, and ordinary microphones are used to perceive the user's voice.

In the digital processing of the sensor signal, we pass the signal to the host computer through serial communication. We obtain the number information of the photosensitive blocked, the value information of the pressure sensor and the information of the sound sensor. Mark them as *pho*, *pes*, *sou*. And form them into one triad $be = (pho, pes, sou)$.

In the word processing of speech signals, We use the Baidu speech recognition to convert the speech into text. Baidu speech recognition is a cloud computing method that puts the burden of calculation and storage onto the cloud, so it can reduce the cost of embedded device development. Besides, it provides a voice client terminal system, internal integrated audio processing, audio codec module, and a complete API [33]. The user can call the voice function service for different scenarios through the combination interface. Hence, we use this intelligent speech recognition method to upload audio-encoded voice data to the cloud, and then output the

final text data through data analysis[35]. We record the perceived speech result as *sp*. Fig. 4 shows the voice perception through Baidu speech recognition.

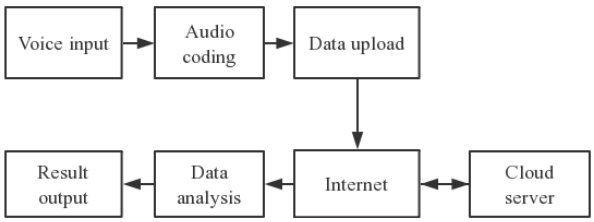


FIGURE 4. Baidu speech recognition process.

After obtain sensor information $be = (pho, pes, sou)$ and voice information *sp*, we define a function definition for identifying two modal information *yb* and *ys*. They are shown as (3) and (4) respectively.

$$\begin{cases} yb = (yph, ype, yso) \\ yph = fb_1\,(xph) = Pho\_xph \cap pho \\ ype = fb_2\,(xpe) = Pes\_xpe \cap pes \\ yso = fb_3\,(xso) = Sou\_xso \cap sou \end{cases} \tag{3}$$

In (3), *yph*, *ype*, *yso* are used to compare the sensory information generated by user behavior and the operational behavior database (TABLE 3), so as to get the identified user operation behavior information $be\_IN = yb$. Among them, $xph = \{1, 2, 3, 4\}$, $xpe = \{1\}$, $xso = \{1, 2\}$.

$$ys = fs\,(xs) = Sp\_xs \cap sp \tag{4}$$

In (4), *ys* is used to compare the voice information sent by the user with the keywords speech database (TABLE 4) to obtain the recognized user voice information $sp\_IN = ys$.

D. MULTIMODAL FUSION UNDERSTANDING MODAL

1) MULTIMODAL FUSION ALGORITHM

In the dual-mode system, the information fusion process can be divided into three levels: data layer fusion, feature layer fusion and decision-making layer fusion [37]. According to the characteristics of this virtual chemical experiment simulation system, decision-making layer fusion is selected. First, each modal matching and recognition process is performed separately. Then, the output or decision of the model is integrated to produce the final decision result, and the decision-level fusion is completed 错误!未找到引用源。.

After perceiving and identifying the information of the two modalities, we propose a multimodal fusion understanding (MMFU) algorithm. *be_IN* and *sp_IN* are recorded as input information for MMFU.

In the algorithm, we first use the 0 and 1 coding method for the recognition results under each label in the two modal databases to form the corresponding sensor label *IBL* and voice label *ISL*. The calculation formula of *IBL* is (5).

$$\begin{cases} IBL = (IB_1, IB_2, IB_3, IB_4, IB_5, 0, 0)^{\mathrm{T}} \\ IB_j = \underset{labelb=j, i=1,2,\ldots,7}{\cup} Ib_i \end{cases} \tag{5}$$

In the solution of *IBL*, its value is obtained by Ib_i $(i = 1, 2, \ldots, 7)$, the calculation formula of Ib_i is (6).

$$Ib_i = \begin{cases} yph|_{xph=i} & i = 1, 2, 3, 4 \\ ype|_{xpe=1} & i = 5 \\ yso|_{xso=i-5} & i = 6, 7 \end{cases} \tag{6}$$

The calculation formula of *ISL* is (7).

$$\begin{cases} ISL = (IS_1, IS_2, IS_3, IS_4, IS_5, IS_6, IS_7)^{\mathrm{T}} \\ IS_j = \underset{labelb=j, i=1,2,\ldots,7}{\cup} Is_i \\ Is_i = ys|_{xs=i} \end{cases} \tag{7}$$

Then, we do mod operation, multiplication and intersection operations on the labels of the two modal information. And use variables *ab*, *as*, *a* and *A* to represent respectively. The calculation formula is as follows.

$$\begin{cases} ab = \|IBL\| \\ as = \|ISL\| \\ a = IBL \cdot ISL \\ A = IBL^{\mathrm{T}} \cap ISL^{\mathrm{T}} \end{cases} \tag{8}$$

Then, we get the key operator λ used to determine the user's intention through the numerical calculation of *A*. The calculation formula is as follows.

$$\lambda = \arg\max_{i}\,(A) \tag{9}$$

The meaning of λ is the value of *i* corresponding to the maximum value of the element in *A*. Among them, $i = 1, 2, \ldots, 7$.

The following is the specific process of the MMFU algorithm. Users' intention can be obtained through MMFU algorithm.

The following example to verify the algorithm. If the intent information set obtained by speech recognition is {transfer, pour water}. The set of intent information recognized by the sensor is {transfer, pour}. After calculation, the corresponding label can be obtained *IBL* and *ISL*. The values are [1 1 0 0 0 0 0] and [1 0 1 0 0 0 0]. The calculated value of *A* is [1 0 0 0 0 0 0]. At this time, the value of λ is 1. That is, the intention after multimodal fusion is dumping.

Another example is below. If the intent information set obtained by speech recognition is {pour, pour water}. The set of intent information recognized by the sensor is {take the appropriate amount of sodium}. The values of *IBL* and *ISL* are [1 1 0 0 0 0 0] and [0 0 0 0 1 0 0]. The values of *A* is [0 0 0 0 0... 0]. At this time, the value of λ is 0. That is, the user's intention is not obtained after fusion. The reason is that there are differences in the intentions represented by the information of the two modalities, which causes the user to act again to confirm the intention.

2) NAVIGATIONAL INTERACTION STRATEGY BASE ON MMFU

Based on MMFU algorithm, we design a navigation interaction strategy. Through this strategy, the effect of the experimental reaction can be visually presented to the user

Algorithm 1 MMFU Algorithm

Input: be_IN, sp_IN;
Output: I;
1: Through the perception and recognition of multimodal information, behavioral intention information be_IN and voice intention information sp_IN are obtained. And express them as $Multi\_IN = (be\_IN, sp\_IN)$;
2:**while** $Multi\_IN$ is not empty **do**
3: Calculate and match $Multi\_IN$ according to formula (5), (6), (7) with the values of the operational behavior database and the keywords speech database to obtain IBL and ISL;
4: Calculate ab, as and a according to formula (8);
5: **if** $ab \neq 0$ or $as \neq 0$ **then**
6: **if** $a \neq 0$ **then**
7: Calculate the key operator λ according to formula (9), $I\_Data \leftarrow \lambda$;
8: **else**
9: $I\_Data \leftarrow \arg\max_i \left(ISL^{\mathrm{T}}\right)$;
10: **end if**
11: **else**
12: $I\_Data \leftarrow 0$;
13: **end if**
14: **for** $i = 1$ **to** 7
15: **if** $i = I\_Data$ **then**
16: $I_i \leftarrow 1$;
17: **else**
18: $I_i \leftarrow 0$;
19: **end if**
20: **end for**
21: Get $I = (I_1, I_2, I_3, I_4, I_5, I_6, I_7)$;
22: **end while**

according to the user's intention, and the user will be given corresponding voice prompts to guide the user in the next operation. In addition, when the user's operation does not match the correct experimental procedure, the user will be given necessary visual and auditory reminders. Figure 5 shows the specific process of the navigation interaction strategy.

In Figure 5, $I = (I_1, I_2, I_3, I_4, I_5, I_6, I_7)$ represent users' intent information obtained by the MMFU algorithm. The intended meanings represented in this paper are ''pour'', ''take out the solid substance'', ''drop the liquid substance'', ''slowly stir or shake'', ''quickly stir or shake'', ''confirm'', ''complete''. The value of sn_1, sn_2 and sw are obtained from the mobile phone's Android terminal (introduced in section 3.1). They respectively represent the name of the substance in beaker 1, the name of the substance in beaker 2 and the weight of the substance in beaker 1 involved in the experiment. w is used to judge the size of the experimental solid substance, which is the threshold value of the weight of the experimental material. This paper set w = 5. $Nag_{ij} = \left(Nv_{ij}, Ns_{ij}\right)$ means navigation given by the system. Nv_{ij}, Ns_{ij} represents visual presentation and voice prompts respectively. When $i = 1, j = 1, 2, \ldots, 9$, Nag_{ij} is the navigation of the user's correct experimental behavior. When $i = 2, j = 1, 2$, Nag_{ij} is the navigation of the user's wrong experimental behavior. The correct and wrong here refer to whether the user's behavior in the experiment conforms to the experimental steps required by the syllabus in real teaching. Conformity is correct, and nonconformity is wrong.

In navigation interaction paradigm based on multimodal fusion understanding, visual presentation uses information enhancement to show the effects of chemical reactions in a virtual scene. According to the experimental requirements, when the user's experimental behavior is correct, the visual presentation makes the reaction phenomenon more obvious than when the experiment is performed in the real world, which can deepen the user's impression. When the user's experimental behavior is incorrect, some dangerous reaction phenomena (such as: corrosion, explosion, etc.) can be presented, which can ensure the safety and make the user deeply understand the incentive mechanism of the reaction.

V. EXPERIMENTAL AND EVALUATION

Based on the analysis of the functions and requirements of VR systems, this paper designs and develops an interactive VR system based on sensing devices. Beaker physical model as shown in Fig. 6. We use Autodesk 3d Max to model the virtual scenes and complete the animation on the Unity3D platform. Based on two experiments (dilution with concentrated sulfuric acid and the reaction of sodium with water), a chemistry experiment platform was built. When the user completes the operation, the system will provide feedback to the user based on visual and auditory cues. The virtual experiment platform built in this paper is shown in Fig. 7.

The test system is designed two complete chemistry experiments, namely, a sodium and water reaction and concentrated sulfuric acid dilution.

A. EXPERIMENTAL SCENE TEST

1) SCENE TEST OF SODIUM AND WATER REACTION

First, the sodium and water reaction experiments were tested. Before the experimental operation, when the user determines the experimental operation of sodium and water reaction by voice input and picks up beaker 2 pours the liquid into beaker 1, the system calculates the user behavior through the multimodal fusion algorithm proposed above. The result of the operation is pouring water, and the feedback of the output is performed through the animation of the experimental scene. The output result is shown in Fig. 8.

Next, the user selects the amount of sodium to be placed by mobile phone, there are some indicators on the mobile phone about the experimental quantification settings. You can choose the weight of sodium, the volume of water, and the selection of various reagents. Then, places the tweezers on top of beaker 1. At this time, the user's behavior is sensed through the light sensor set on the beaker and the data input by the mobile phone. When the user chooses the proper amount

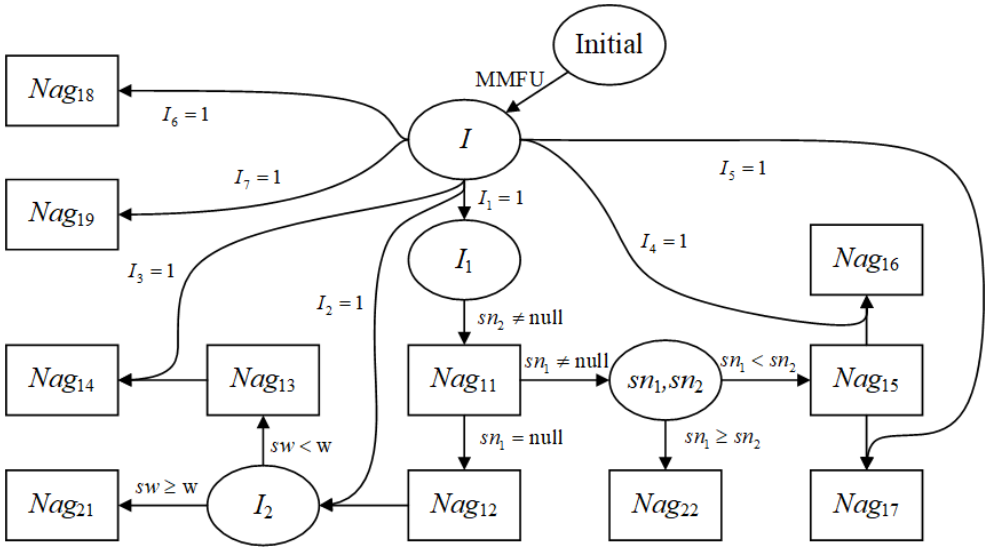


FIGURE 5. Navigation interaction strategy.

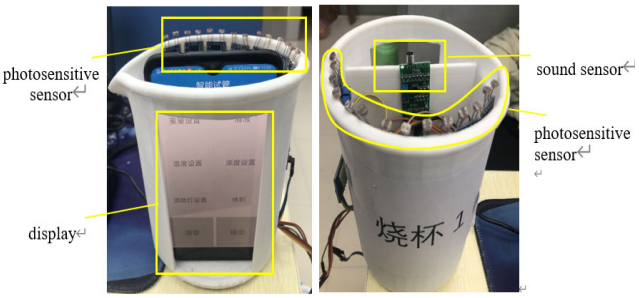


FIGURE 6. Beaker platform.

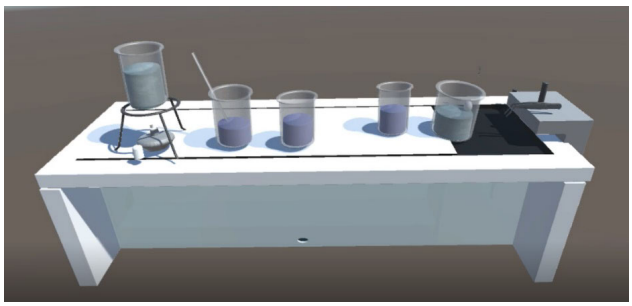


FIGURE 7. Virtual experiment platform.

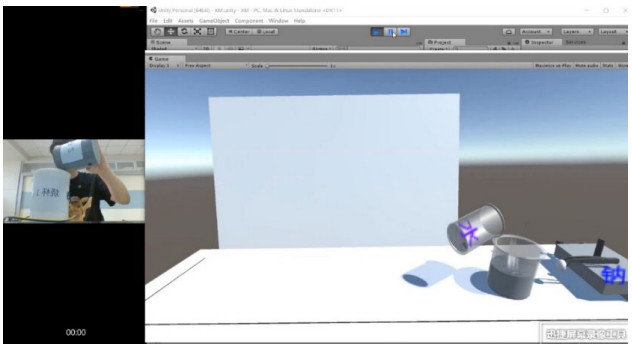


FIGURE 8. Sodium/water reaction experiment system.

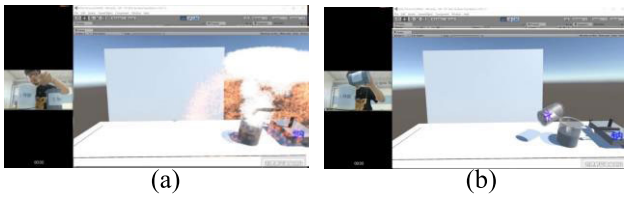


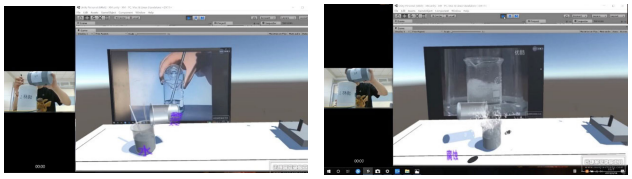
FIGURE 9. The sodium/water reaction experiment. (a) Specific operation interface. (b) Error operation result.

of sodium, we can see the animation of the sodium and water reaction when a small amount of sodium is put in the beaker, a real video and an audio explanation of the reaction between sodium and water, and the problem description. When a large amount of sodium is put in the beaker, the system first tells the user that the operation is dangerous. If the user wants to continue the experiment, the system will output an explosion reaction phenomenon. The specific operation interface and error operation result is shown in Fig. 9.

2) SCENE TEST OF CONCENTRATED SULFURIC ACID DILUTION

First, the user can select the concentrated sulfuric acid dilution experiment by voice input at the beginning stage; then, the user needs to select the reagents in beaker 1 and beaker 2 through the mobile phone; finally, the user picks up beaker 2 and pours the liquid into beaker 1. If the user chooses to pour concentrated sulfuric acid into the water, the operation matches the correct step for diluting concentrated sulfuric acid. At this time, the system will output an animation of concentrated sulfuric acid dilution and a real video. The user

can control the stirring speed of the glass rod by the voice input commands of quick stirring and slow stirring. The specific operation interface is shown in Fig. 12 (a). If the user chooses to pour water into concentrated sulfuric acid, the system will auditorily tell the user that the operation is dangerous. If the user wants to continue the operation, the system will output the animation of the concentrated sulfuric acid splashing onto the desktop, and the desktop will be corroded by concentrated sulfuric acid. The error operation result is shown in Fig. 10 (b).



(a) (b)

FIGURE 10. The concentrated sulfuric acid dilution experiment. (a) Specific operation interface. (b) Error operation result.

When the user is using the device, the user's hand is operating the experimental instrument, but the eyes always have to look at the real screen, which is a great inconvenience to the user. To solve this hand-eye consistency problem, we use a beaker. A mobile phone is set on the computer screen, and the virtual screen is displayed synchronously with the computer screen so that the user can watch the beaker to complete the whole experiment. The actual operation is shown in Fig. 11.

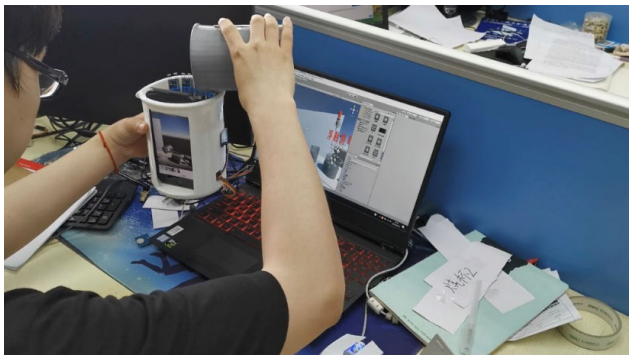


FIGURE 11. Sodium/water reaction experiment system.

B. EXPERIMENTAL RESULTS AND EVALUATION

1) VERIFICATION OF MULTIMODAL FUSION UNDERSTANDING ALGORITHM

We designed the following experiment to verify the multimodal fusion understanding algorithm proposed in this paper to determine whether the user's intention can be better understood. We organized 20 operators with chemistry learning experience to conduct the experiment on the reaction of Na with H2O. The experimental requirements are as follows:

1. Each operator uses the single-modal, multimodal fusion, multimodal fusion understanding method to complete the experiment in turn in the noise-free environment, small laboratory environment and normal classroom environment;

2. Each operator completes the experiment at one time, and the interval between each step should not be too long.

3. Each operator conducts human-machine dialogue according to his normal response speed, and makes correct adjustments according to his normal response speed under navigation prompts.

During the operation, the recorder needs to record the number of successful times of 20 operators by use the single-modal, multimodal fusion, multimodal fusion understanding method in the different environment. Fig. 12 shows the results.

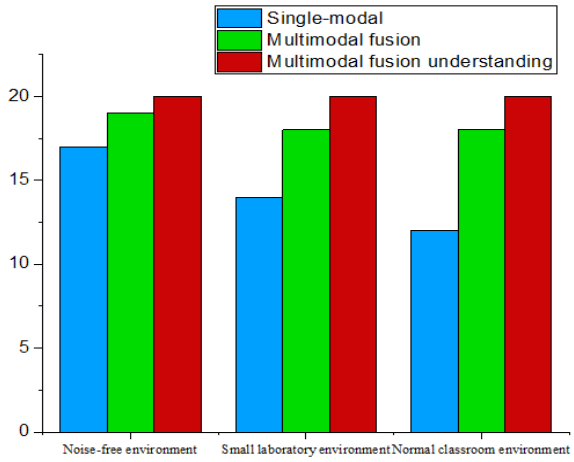


FIGURE 12. The number of successful times.

As we can see in Fig. 12, the number of successful times of 20 operators using single-modal, multi-modal and multimodal fusion understanding method in the noise-free environment is not much different, being 17, 19 and 20 respectively. However, the single-modal in the noise environment shows obvious instability, and the correct rate is below 70%. The correct rate of multimodal fusion is almost the same in the small laboratory environment and normal classroom environment, which shows the good stability. No matter what the environment, the correct rate of multimodal fusion understanding is 100%, which shows that the multimodal fusion understanding model allows users to accurately complete the experimental steps according to the navigation prompts, so as to achieve the best teaching effect.

2) CONPARATIVE EXPERIMENT

We compare the teaching effect and user experience of the NOBOOK [38] experiment and our multimodal fusion understanding experiment. The NOBOOK experiment uses the mouse as the input source, and moves the virtual object by clicking the mouse. It only conducts experiments in a single-channel way. But our multimodal fusion understanding experiment adopts the dual-channel and navigational interaction paradigm of sensor and voice, which frees the

FIGURE 13. NOBOOK experimental system.

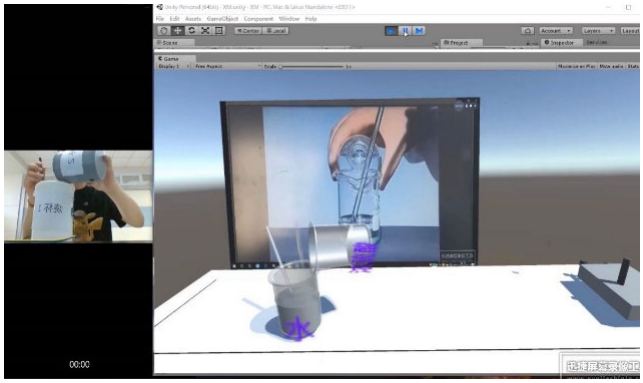


FIGURE 14. Intelligent chemistry experiment simulation system.

user from the control of the mouse, thereby realizing natural virtual intelligent chemical experiment interaction.

During the experiment, 10 experienced teachers and 50 students from different middle schools are invited to conduct comparative experiments. Among them, the students are between 12 and 18 years old. First, all experimenters are asked to conduct two experiments in turn. Then, all experimenters need to comprehensively score the evaluation indicators of the two experiments based on their own experience.

From the perspective of teacher teaching, we evaluate the teaching effect, experience authenticity, effect authenticity, repeatability exploration ability and safety of the NOBOOK experiment and our experiment. The six indicators are named as I1 to I6. There are 6 questions in total and each question is a statement. The 10 teachers are asked to use a five-point scale to rate the degree of agreement with the question, the degree of 0-1 indicates strongly disagree, the degree of 1-2 indicates disagree, the degree of 2-3 indicates neutral, the degree of 3-4 indicates agree, and the degree of 4-5 indicates strongly agree. Fig. xx shows the evaluation result of teaching effect.

From Fig. 16, we can see that the overall evaluation of our multimodal fusion understanding experiment is higher than the NOBOOK experiment. In the I1, the score difference of the two experiments is the largest, with a difference of 2.540, which indicates that the multimodal fusion algorithm based on speech and sensing greatly reduces the user's memory load and meet the usability of teaching. In the I3 and I5, the evaluation of our multimodal fusion understanding

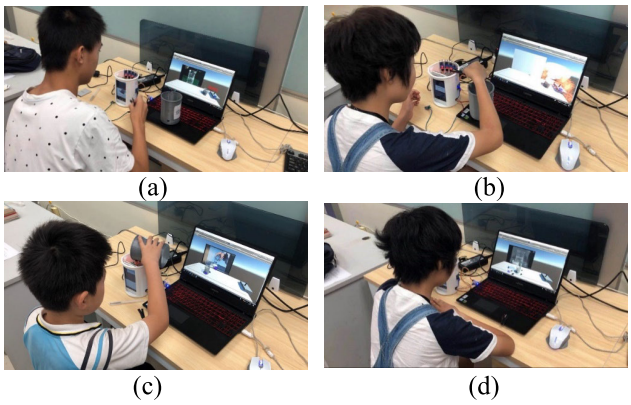


FIGURE 15. Students experimental scene. (a) Sodium reacting with water. (b) An error in the reaction of sodium with water. (c) Diluting concentrated sulfuric acid. (d) An error in the dilution of concentrated sulfuric acid.

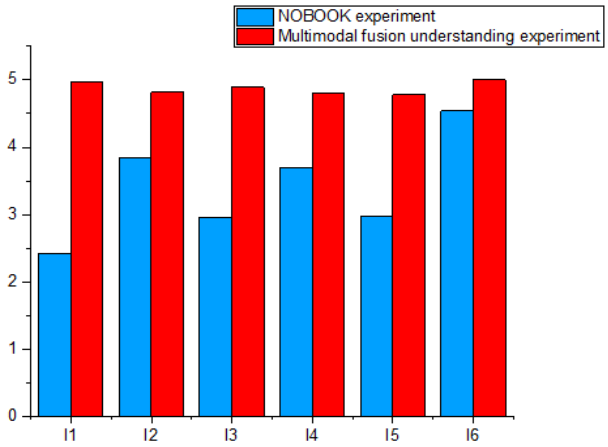


FIGURE 16. Comparison chart of the system function evaluation scores.

experiment is significantly higher than the NOBOOK, which is 65% and 60% higher. It shows that the navigation interaction paradigm based on the intent understanding model greatly improves the accuracy of the experiment. At the same time, the intelligent beaker experimental device greatly increases the immersion of the students in the experiment and improves the authenticity of the experimental results.

From the perspective of students, we use the NASA evaluation index to evaluate their user experience. NASA is divided into mental demand (MD), physical demand (PD), time demand (TD), effort (E), performance (P) and frustration (F). There are 6 questions in total. The NASA evaluation index also uses a five-point scale, in which 0-1 indicates that the cognitive burden is small, the 1-2 indicates that the cognitive burden is relatively small, 2-3 indicates that the cognitive burden is neutral, 3-4 indicates that the cognitive burden is relatively large, and 4-5 indicates that the cognitive burden is large. Fig. 17 shows user experience evaluation results.

For our multimodal fusion understanding experiment, the PD score is the lowest at 1.920, and the scores of MD, TD, P and F are relatively low. But the E score is the highest

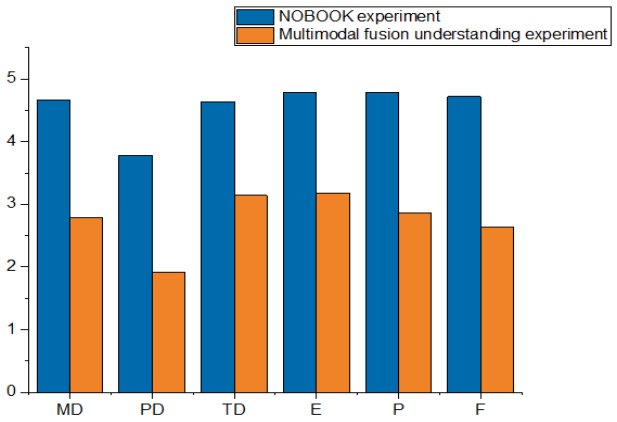


FIGURE 17. Comparison chart of the user experience evaluation score.

at 3.180 which shows that it is difficult for students to do the experimental operation in the process of the experiment. Therefore, the difficulty of the student's operation needs to be reduced in the experimental operation design. Compared with NOBOOK, the average value of each indicator score of the user experience in our experiment is lower than the average value of NOBOOK, indicating that the user evaluation of our experiment is higher and the user acceptance is stronger.

VI. CONCLUSION

The virtual chemistry experiment simulation system has the potential to change the teaching strategies. Students no longer have to memorize chemical equations and experimental reactions but instead use experiments to better understand the subject. The system increases students' interest in learning and their ability to use knowledge by developing the students' own skills. Moreover, conducting experiments autonomously can help nurture students' exploration spirits and cultivate students' innovative thinking and practical abilities. More importantly, the virtual chemistry experiment simulation system greatly reduces the consumption of experimental materials and does not need to address waste left after the experiment, which is very convenient and environmentally friendly.

Compared with the traditional virtual simulation experiment system. The intelligent navigation chemical laboratory scheme based on multi-modal fusion proposed in this paper does not require a keyboard and mouse, allowing the user to rely on his or her hand to operate the intelligent experimental beaker with voice commands, which greatly increases the immersion of the students in the experiment. In addition, we use a multimodal fusion (MMFU) algorithm based on decision-making layer fusion to achieve dual-channel information fusion from speech and sensing. By combining the navigation interaction paradigm based on the intent-understanding model, the success rate of the experimental operation is greatly improved. Compared with the NOBOOK system, students no longer need to memorize the operation steps and do not use memorized gesture commands, which greatly reduces the user's memory load, improves the students' experimental interest, and deepens the students' understanding of the theoretical concepts. Therefore, better experimental teaching effects and user evaluations are obtained.

At the same time, this article also has some areas for improvement. The photosensitive sensor on the designed smart beaker will be affected by light to a certain extent, and the next step will be to improve this problem algorithmically. In addition, some particle effects in the visual presentation of the virtual scene are not realistic enough, and the next step will continue to optimize to make the expression more realistic.

REFERENCES

[1] W. Zhang *et al.*, "Exploration of the construction of environmental engineering virtual simulation experiment center based on OBE mode," (in Chinese), *Exp. Technol. Manage.*, vol. 36, no. 269(01), pp. 270–273, Jan. 2019.

[2] D. Sprute, K. Tönnies, and M. König, "Virtual border teaching using a network robot system," 2019, *arXiv:1902.06997*. [Online]. Available: http://arxiv.org/abs/1902.06997

[3] B. Peixoto, D. Pinto, A. Krassmann, M. Melo, L. Cabral, and M. Bessa, "Using virtual reality tools for teaching foreign languages," in *Proc. WorldCIST Conf., Adv. Intell. Syst. Comput.*, vol. 932, 2019, pp. 581–589.

[4] Y.-C. Yu, "Teaching with a dual-channel classroom feedback system in the digital classroom environment," *IEEE Trans. Learn. Technol.*, vol. 10, no. 3, pp. 391–402, Jul. 2017.

[5] Z. Wang, Q. Liu, H. Yao, and J. Chen, "Virtual chime-bells experimental system based on multi-modal fusion," in *Proc. Int. Conf. Educ. Innov. Technol. (EITT)*, Wuhan, China, Oct. 2015, pp. 64–67.

[6] S. Park and J. Ryu, "Exploring preservice teachers' emotional experiences in an immersive virtual teaching simulation through facial expression recognition," *Int. J. Hum.–Comput. Interact.*, vol. 35, no. 6, pp. 521–533, Apr. 2019.

[7] I. O. Adam, J. Effah, and R. Boateng, "Activity theory analysis of the virtualisation of teaching and teaching environment in a developing country University," *Edu. Inf. Technol.*, vol. 24, no. 1, pp. 251–276, Jan. 2019.

[8] D. Liu, X. Liu, W. Zhang, A. Lu, and R. Huang, "The potential, progress and challenge of virtual reality technology education application," (in Chinese), *Open Edu. Res.*, vol. 22, no. 4, pp. 25–31, 2016.

[9] X. Yang, X. Li, and W. Xing, "Construction and trend analysis of teaching big data practice framework for wisdom education," (in Chinese), *Res. Electron Edu.*, vol. 39, no. 10, pp. 21–26, 2018.

[10] H. S. Al-Khalifa, "CHEMOTION: A gesture based chemistry virtual laboratory with leap motion," *Comput. Appl. Eng. Edu.*, vol. 25, no. 6, pp. 961–976, Nov. 2017, doi: 10.1002/cae.21848.

[11] A. Ables, "Augmented and virtual reality: Discovering their uses in natural science classrooms and beyond," in *Proc. ACM SIGUCCS Annu. Conf.*, Seattle WA, USA, Oct. 2017, pp. 61–65.

[12] L. Ma, B. Ji, Q. Huang, J. Duan, P. Yang, H. Wang, and W. Zhao, "Intelligent equipment for photocatalytic decomposition of ethylene and simulation design," *IFAC-PapersOnLine*, vol. 51, no. 17, pp. 503–508, 2018, doi: 10.1016/j.ifacol.2018.08.159.

[13] I. Gherguleşcu, A. N. Moldovan, C. H. Muntean, and G. M. Muntean, "Interactive personalised stem virtual lab based on self-directed learning and self-efficacy," in *Proc. Adjunct Publication 27th Conf. User Modeling*, New York, NY, USA, Jun. 2019, pp. 355–358.

[14] S. Munawar, S. K. Toor, M. Aslam, and M. Hamid, "Move to smart learning environment: Exploratory research of challenges in computer laboratory and design intelligent virtual laboratory for eLearning technology," *EURASIA J. Math., Sci. Technol. Edu.*, vol. 14, no. 5, pp. 1645–1662, Feb. 2018.

[15] Z. Wu, L. Cai, and H. M. Meng, "Multi-level fusion of audio and visual features for speaker identification," in *Proc. Int. Conf. Biometrics*, Piscataway, NJ, USA, 2006, pp. 493–499.

[16] S. Munawar, S. K. Toor, M. Aslam, and M. Hamid, "Move to smart learning environment: Exploratory research of challenges in computer laboratory and design intelligent virtual laboratory for elearning technology," *EURASIA J. Math., Sci. Technol. Edu.*, vol. 14, no. 5, pp. 1645–1662, Feb. 2018.

[17] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1–15, Nov. 2002.
[18] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia (MULTIMEDIA)*, New York, NY, USA, 2005, pp. 399–402.
[19] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.
[20] S. Qi, V. D. Calhoun, T. G. M. van Erp, J. Bustillo, E. Damaraju, J. A. Turner, Y. Du, J. Yang, J. Chen, Q. Yu, D. H. Mathalon, J. M. Ford, J. Voyvodic, B. A. Mueller, A. Belger, S. McEwen, S. G. Potkin, A. Preda, T. Jiang, and J. Sui, "Multimodal fusion with reference: Searching for joint neuromarkers of working memory deficits in schizophrenia," *IEEE Trans. Med. Imag.*, vol. 37, no. 1, pp. 93–105, Jan. 2018.
[21] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Inf. Fusion*, vol. 51, pp. 259–270, Nov. 2019.
[22] L. Guo, "Latent variable algorithms for multimodal learning and sensor fusion," 2019, *arXiv:1904.10450*. [Online]. Available: http://arxiv.org/abs/1904.10450
[23] J. Mi, S. Tang, Z. Deng, M. Goerner, and J. Zhang, "Object affordance based multimodal fusion for natural human-robot interaction," *Cognit. Syst. Res.*, vol. 54, pp. 128–137, May 2019.
[24] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60736–60751, 2019.
[25] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning," *Sensors*, vol. 19, no. 7, p. 1716, Apr. 2019.
[26] C. Yang, D. Wang, Y. Zeng, Y. Yue, and P. Siritanawan, "Knowledge-based multimodal information fusion for role recognition and situation assessment by using mobile robot," *Inf. Fusion*, vol. 50, pp. 126–138, Oct. 2019.
[27] R. M. Jiang, A. H. Sadka, and D. Crookes, "Multimodal biometric human recognition for perceptual human–computer interaction," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*, vol. 40, no. 6, pp. 676–681, Nov. 2010.
[28] V. Peruffo Minotto, C. Rosito Jung, and B. Lee, "Multimodal multichannel on-line speaker diarization using sensor fusion through SVM," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1694–1705, Oct. 2015.
[29] Z.-T. Liu, F.-F. Pan, M. Wu, W.-H. Cao, L.-F. Chen, J.-P. Xu, R. Zhang, and M.-T. Zhou, "A multimodal emotional communication based humans-robots interaction system," in *Proc. 35th Chin. Control Conf. (CCC)*, Chengdu, China, Jul. 2016, pp. 6363–6368, doi: 10.1109/ChiCC.2016.7554357.
[30] M. Ghayoumi, M. Thafar, and A. K. Bansal, "Towards formal multimodal analysis of emotions for affective computing," in *Proc. DMS*, 2016, pp. 48–54.
[31] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: ACL Press, 2014, pp. 1724–1734.
[32] F. Peng, X. Yan, H. Wang, J. Zhong, and J. Pu, "Efficient database access mechanism based on Android mobile phone," (in Chinese), *Comput. Eng. Des.*, vol. 34, no. 12, pp. 4109–4113, 2013.
[33] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: ACL Press, 2015, pp. 1412–1421.
[34] D. Xu, S. Tao, and Y. Liu, "Research and practice of electronic medical records based on voice cloud," (in Chinese), *China Digit. Med.*, vol. 7, no. 3, pp. 15–18. 2012.
[35] F. Lin, Y. Luo, F. Kong, Y. Zhang, Y. Qiao, and Y. Qian, "A design of intelligent robot voice interaction system based on cloud platform," (in Chinese), *Electron. Test*, no. 3 pp. 40–42, 2018.
[36] L. Barker and J. Gruning, "The student prompt: Student feedback and change in teaching practices in postsecondary computer science," in *Proc. IEEE Frontiers Edu. Conf. (FIE)*, Oct. 2014, pp. 1–8.
[37] J. Wang, L. Ci, and K. Yao, "A survey of feature selection methods," (in Chinese), *Comput. Eng. Sci.*, vol. 27, no. 12, pp. 72–75, 2005.
[38] Nobook System. [Online]. Available: https://www.nobook.com/huaxue.html

DI DONG received the B.S. degree from the University of Jinan, in 2018, where he is currently pursuing the master's degree. His research interest includes intelligent information processing.

ZHIQUAN FENG (Member, IEEE) received the master's degree from Northwestern Polytechnical University, China, in 1995, and the Ph.D. degree from the Department of Computer Science and Engineering, Shandong University, in 2006. He is currently a Professor with the School of Information Science and Engineering, University of Jinan. He has published more than 50 articles on international journals, national journals, and conferences in recent years. His research interests include human hand tracking/recognition/interaction, virtual reality, human–computer interaction, and image processing.

JIE YUAN is currently pursuing the master's degree in computer science and information technology with the University of Jinan. His research interests include multimodal fusion and human–computer interaction.

XIN MENG is currently pursuing the master's degree with the School of Information Science and Engineering, University of Jinan. Her research interests include multimodal fusion and human–computer interaction.

JUNHONG MENG is currently pursuing the master's degree with the School of Information Science and Engineering, University of Jinan. Her research interests include human–computer interaction and intelligent information processing.

DAN KONG is currently pursuing the master's degree with the School of Computer Science and Information Technology, University of Jinan. Her research interests include VR and human–computer interaction.