Smart Beaker Based on Multimodal Fusion and Intentional Understanding

Di Dong  
Department of Information Science and Engineering,  
University of Jinan, Jinan 250022, China  
Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China  
2113886351@qq.com

Zhiquan Feng *  
Department of Information Science and Engineering,  
University of Jinan, Jinan 250022, China  
Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China  
ise_fengzq@ujn.edu.cn

Jinglan Tian  
Department of Information Science and Engineering,  
University of Jinan, Jinan 250022, China  
Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China  
ise_tianjl@ujn.edu.cn

ABSTRACT

In the current simulation experiment system, the experimental design of single mode is less interactive and less accurate. In order to solve this problem, this paper proposes an experimental interaction kit based on sound and sensor, and designs a multimodal fusion and intent understanding algorithm. Firstly, the method of multi-sensor signal extraction and speech feature extraction is introduced. Then, based on the results obtained by the two methods, an algorithm based on decision-level fusion is studied, which solves the problem of perception of user's operation intention in virtual chemistry experiments. Finally, the usability of the multimodal intent understanding algorithm proposed in this paper is verified by designing a complete chemical experiment system. Experiments show that the multimodal intent understanding algorithm based on sensor and speech input is due to a single modality in terms of interactivity and accuracy, and the physical interaction suite designed in this paper greatly improves the intelligence and interactivity of the system.

CCS Concepts

• Information Human-centered computing→Human-computer interaction.

Keywords

Multimodal fusion; Understanding intention; Chemical simulation experiment; Smart beaker

1. INTRODUCTION

The reaction between sodium and water is one of the important experiments in the middle school curriculum and an important experiment in the introduction of chemical experiments. However, due to the high cost of sodium metal and the complexity and danger of experimentation, it is not possible for each student to experience the experimental process steps. The lack of interaction between teachers and students in traditional experimental teaching is not conducive to teachers mastering the learning situation of each student. If the wrong operation steps are not able to make timely prompts, it is difficult to guide each student one by one.

In order to solve these problems, this paper proposes a teaching-oriented experimental device based on multi-modal fusion. The main work of this paper is as follows: First, it solves the problem of lack of experimental items in primary and secondary schools, and solves the problem of lack of teacher guidance in the teaching process by using experimental navigation and experimental result scoring system. Second, the whole process steps and experimental phenomena of sodium and water experiments are given intelligent features, which enhance the interaction between students and equipment. Third, the use of virtual and real fusion technology to explain the experimental phenomena and reaction principles clearly, making the experimental phenomenon more obvious, students' understanding of the phenomenon is more profound.

2. RELATED WORK

In the field of virtual experiment teaching, Liu Dejian et al. (2016) [1] studied the potential, application and challenges of VR technology in education based on the analysis of virtual reality contextual situational cognition (VRSC Model). Xu et al. (2018) [2] discussed the effective application of virtual simulation experiments in physics experiment teaching under finite conditions. It is pointed out that the main problems in real physics experiments are the difficulty of practical operation, the limitation of real experiments by the teaching environment, and abstraction. The teaching content makes it difficult for students to master key knowledge points, and virtual teaching experiments can make up for these shortcomings. Barbara et al. (2018) [3] studied the role of the Second Life Online Virtual Environment (SL) in teaching and found that SL can be used as an innovative way to learn research skills and improve the learning experience of students. Peixoto et al. (2019) [4] 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. Experiments have shown that this method is not only attractive, but also helps to motivate students.

Based on the above research, we can find the virtual experiment teaching can make up for the shortcomings of traditional experimental teaching and stimulate students' motivation to study, which plays an increasingly important role in today's classroom. However, the research on VR technology in the field of teaching is still not comprehensive, and there are few studies on applying VR technology to the reality of chemistry teaching. Therefore, a virtual experimental device for chemistry teaching is proposed. In
order to improve the immersion and accuracy of user operations, this paper uses a multi-modal fusion algorithm based on sensing and speech. In addition, this paper uses the experimental navigation system to solve the problem of lack of teacher guidance in the teaching process.

3. INTELLIGENT BEAKER DESIGN FOR MULTIMODAL INTERACTION

The design of the intelligent beaker is the basis of the entire chemical experiment. We need to use the beaker to sense the user's behavior, including picking up the beaker, moving the beaker, and rotating the beaker to achieve different experimental operations. This requires designing various sensing devices on the beaker to receive user information. At the same time, we have added voice prompts and voice judgments to improve the fault tolerance of the entire experimental environment and make the whole experimental system more intelligent. The design physical map is shown below.

As shown in Figure 1, the beaker physical design includes an infrared range finder and an attitude sensor. The infrared rangefinder is used to sense the position of the beaker and other equipment on the test bench. The attitude sensor is used to implement the detection of the three-dimensional information of the beaker. The position and posture of the beaker are displayed in a three-dimensional scene by unity3D. The touch sensor is used to sense the device and experiment process that the user is operating at this time.

3.1 Multimodal fusion algorithm framework

This paper proposes a new multimodal fusion and intention understanding strategy. First, at the multimodal input layer, user behavior is sensed in real time through sensing devices and voice input. Then, the sensing intent is obtained through the interaction device sensing, and the sensing intention perception is classified into (posture perception, distance perception, touch perception). Moreover, the obtained sensing intent and the intention of the voice behavior are merged to obtain the user's behavior intention. Finally, the corresponding user behavior is output. The overall algorithm framework is shown in Figure 2.

3.2 Beaker gesture perception

The device uses a nine-axis attitude sensor and an infrared ranging sensor to obtain three-dimensional information of the beaker. Users can perform arbitrary rotation and movement operations on our beakers. The attitude sensor and infrared sensor will send the data to the computer through the serial communication, and the real-time and accurate information feedback will be given in the virtual scene.

The data transmitted by the sensor is the angular velocity, acceleration and magnetic induction intensity measured on behalf of the three axes. It needs to be solved into the attitude angle information. The elevation angle \( \rho \) (the angle between the X axis and the ground) and the roll angle \( \varphi \) (the Y axis angle to the ground) and heading angle \( \alpha \) (angle of rotation about the Z axis). By calculating the acceleration components of the three axes, the angular attitude of the three angles can be derived to achieve three-dimensional rotation of the cup. The algorithm is as follows:

Algorithm 1: Beaker trajectory sensing algorithm

**Input:** Enter the attitude sensor acceleration \( A_x, A_y, A_z \);

**Output:** The beaker track.

1: Enter the acceleration components \( A_x, A_y, A_z \) of the three axes X, Y, Z;

2: Using the trigonometric function relationship to solve the angular pose, the formula is:

\[
\begin{align*}
\rho & = \arctan \left( \frac{A_y}{\sqrt{A_x^2 + A_z^2}} \right) \\
\varphi & = \arctan \left( \frac{A_x}{\sqrt{A_y^2 + A_z^2}} \right) \\
\alpha & = \arctan \left( \frac{\sqrt{A_x^2 + A_y^2}}{A_z} \right)
\end{align*}
\]

(1)

3: Subtract the obtained \( \rho, \varphi, \alpha \) from the initial three angles \( \rho_1, \varphi_1, \alpha_1 \) of the beaker to obtain the current angular attitude \( \rho_2, \varphi_2, \alpha_2 \);

4: Return the extracted keywords.

We assign the angular poses \( \rho_2, \varphi_2, \alpha_2 \) obtained by the attitude sensor to the pose of the virtual beaker model under unity. Then, by placing the attitude sensor in the real beaker, we can synchronize the beaker in the virtual scene with the attitude of the real beaker, thereby judging the behavior of the user at this time.
by judging the angle at which the beaker is tilted. If the bevel angle of the beaker is greater than a certain pre-value $r$, we believe that the user is performing a pour operation at this time.

3.3 Beaker distance information perception

In this paper, the system places the sensor on the side wall of the beaker. When we pick up the beaker and move it, the infrared distance measuring sensor can sense the distance from the object in front. We can judge the user's behavior intention according to this distance. For example, when the distance between the beaker and the object in front is less than 5 cm, we will prompt the user by voice whether or not to perform the pouring operation.

3.4 Touch perception

Touch-sensing devices are provided on each experimental device, and each touch sensor has a corresponding unique number. The system recognizes the experimental device being used at this time by the user's touch and judges the intention of the user's operation. If the user touches the touch sensor on the dice model at this time, when the computer returns a unique number on the dice sensor, the system understands that the user is using the dice to perform the action of placing sodium. If the user touches the touch sensor on the plastic dropper model at this time, then the system judges that the user's behavior is intended to add phenolphthalein reagent.

4. SENSING AND SPEECH BASED MULTIMODAL FUSION

4.1 Speech intent recognition

Cloud speech recognition technology is a method of recognizing and processing speech in a "cloud computing" manner, which puts the pressure of calculation and storage into the cloud, thereby reducing the cost of embedded device development. Developers can focus more on the needs of the application and reduce the cycle of application development [5]. The cloud speech recognition technology also provides a voice client terminal system, an integrated audio processing and audio codec module, and provides a complete API interface. The user can call the service voice function service of different scenarios through the combined interface [6].

This paper uses Baidu cloud speech recognition technology. Firstly, through the intelligent speech recognition algorithm, the input speech is recognized as text and text. Then, the lexical analysis function of Baidu speech is used to analyze the user's instructions, and the word set of the user instruction is obtained. Finally, the keywords in the Italian gallery are matched to obtain the final recognition result. The keyword matching is realized by distance transformation. In this paper, the Euclidean distance transform is selected as the distance metric. The formula is:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$  \hspace{1cm} (2)

Among them, the point $(x_i, y_i)$ represents the word of speech recognition, and the point $(x_j, y_j)$ represents the keyword in the database, which is the Euclidean distance between the two points. The similarity measure is compared with the preset threshold to determine whether the match is successful. If it is 0, the match is successful, otherwise the match fails.

Where point $(x, y)$ represents the word of speech recognition, point $(x_i, y_i)$ represents the keyword in the database, and $d_{ij}$ is the Euclidean distance between the two points. Compare $d_{ij}$ as a similarity measure with a preset threshold to determine whether the match is successful. If $d_{ij}$ is 0, the match is successful, otherwise the match fails.

Algorithm 2 Speech recognition algorithm

**Input:** Enter the experimental operation voice;

**Output:** The voice intent.

1: Identify the input speech into text by the intelligent speech recognition algorithm of Baidu Cloud Voice;

2: Using the lexical analysis function of Baidu speech to analyze the user's instructions and obtain the word set of the user instruction;

3: Match the word $(x_i, y_i)$ in the word collection to the keyword $(x_j, y_j)$ in the gallery, and the matching process is implemented by the Euclidean distance transformation. The formula is: $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. If $d_{ij} = 0$, the match is successful, and the matched keywords are output, otherwise the match fails.

4: If successful, the keyword is output as a voice intent.

4.2 Decision layer fusion

According to the characteristics of this virtual chemical simulation experiment system, the decision layer fusion is selected. The decision-making layer has strong anti-interference ability and good fault tolerance, and can effectively reflect different types of information on each side of the environment or target [7]. Therefore, after obtaining the sensor perception and speech recognition results respectively, the multi-modal data is transmitted to the intent fusion layer, and the final recognition result is obtained by the decision layer fusion method.

The modal fusion rule adopts the vector multiplication rule: first, input the sensing intention and the voice intention obtained at the intention understanding layer; second, compare the sensing intention and the voice intention with the intention in the user's intention library, and mapped to the corresponding location of the user's gallery; thirdly, it is represented by a 0, 1 encoding in the form of a vector, that is, the position of the element in the Italian library and the intent set is marked with 1, and there is no intention in the gallery. The element position is marked with 0, the sensor intent vector is $X_1$, and the phonetic intent vector is $X_2$. Fourth, the two vectors are multiplied to obtain a new vector $Y$, which is the merged intent vector. The formula is as follows:

$$Y = X_1X_2^T$$ \hspace{1cm} (3)

If $Y = 0$, the modal fusion fails, the sensing intent and the voice intention do not match, and the next step cannot be performed. The system will use the voice to make an error message to the user, and the user is required to re-enter the voice command; if $Y$ is not a zero vector, the modal fusion is successful, the system will remind the user to proceed to the next step and pass the result to the interactive application layer.

Among them, the user's intention library is shown in Table 1:

| Intention1 | Intention2 | Intention3 | Intention4 |
|------------|------------|------------|------------|
| Pour water | Grab sodium | Adding phenolphthalein reagent | Mobile beaker |

Table 1. The user's intention library
For example, when the voice is intended to be "dropping phenolphthalein reagent", $X_1=[0\ 0\ 1\ 0]$. If the sensor senses by touch at this time is "adding phenolphthalein reagent", that is $X_2=[0\ 0\ 1\ 0]$, then $Y=[0\ 0\ 1\ 0]$ and $Y$ are not zero vectors, then the modal fusion is successful; conversely, if the intention acquired by the sensor at this time is not “adding phenolphthalein reagent”, then $Y=0$, the modal fusion fails. The algorithm is as follows:

| Algorithm 3 Multimodal Fusion Algorithm for Sensing and Speech |
|---------------------------------------------------------------|
| **Input:** Enter the sensing intent and voice intent;          |
| **Output:** Fusion result.                                    |
| 1: Enter the sensing intent and voice intent;                 |
| 2: Map the sensing intent and voice intent to the             |
| corresponding location of the user's intention library;       |
| 3: Using 0, 1 encoding to represent them in the form of        |
| vectors, respectively, to obtain the sensing intent vector $X_i$|
| the speech intent vector $X_i$;                               |
| 4: Multiply $X_i$ and $X_i'$ to get the fused intent vector.   |
| The formula is: $Y = X_i X_i'$                               |
| 5: Determine if the fusion is successful. When $Y = 0$, the    |
| fusion fails, identify the type of error and report an error, |
| wait for the next input information, go to step 2, otherwise  |
| go to step 5;                                                 |
| 6. The integration is successful and the final intent is output.|

5. System test

5.1 Performance Testing

This paper designed a complete chemical experiment (sodium and water reaction experiments). The experimental process of sodium and water reaction can be accomplished by a combination of sensor and speech. This article uses Unity3D to complete the construction of the virtual scene, through the combination of animation and particle system to complete the experimental reaction, the construction scenario is shown in Figure 3.

At the beginning of the experiment, the system will give a voice prompt "At this time you are conducting a chemical experiment for sodium and water reaction, please determine whether to start the experiment", then you need to input the voice to determine the start of the experiment. When the user determines that the experiment starts, the system will give a voice prompt "Please follow the experimental steps on the blackboard to complete the experiment", and at this time there will be experimental steps on the blackboard that require user operation, as shown in Figure 4.

Next, we need the user to operate the beaker of our design to complete the experiment. Below we will make the left beaker into No. 1 beaker and the right beaker as No. 2 beaker. First, we will carry out the pouring operation according to the requirements of the experimental procedure. Then we will take the No. 1 beaker on the left side and gradually approach the No. 2 beaker. At this time, the distance sensor on the side wall of the No. 1 beaker will sense the change of the distance between the two beakers and return it to the system. The beaker on the virtual side will also change the corresponding beaker distance. When the distance between the two beakers is close, the system will prompt us whether we want to carry out the operation of pouring water. At this time, we merge the signals of the two channels by voice confirmation and judge the user's intention. If the user confirms the pouring of water and tilts the beaker, the output is shown in Figure 5.

After the completion of the pouring, the experimental step is to add the phenolphthalein solution. At this time, we pick up the plastic dropper and touch the touch sensor. At this time, the voice will prompt “whether or not to add phenolphthalein reagent”, we pass the voice and sensor. The two-modal fusion algorithm is used to further judge the user's intention. Figure 6 below shows the scenario when the phenolphthalein reagent is added. Next, the final step of the experiment is to put the sodium into the water, the user needs to pick up the dice, the same as above, we also judge the user's intention by the user's intention through the multimodal intent understanding algorithm mentioned above. When the system receives the user behavior, if it is determined that sodium is placed, the operation of putting sodium into the corresponding virtual scene is performed. The reaction between sodium and water is shown by animation and video as shown in Figure 7.
5.2 Comparative test of multimode intent

In order to compare the effects of the single-modal and bimodal and intentional understanding models, we selected 20 middle school students to use two models to test the degree of completion of the two tests. The total number of steps in the experiment is 6. If the number of steps completed in the experiment is 3, the degree of completion is 50%. It can be seen from the test results in Figure 8 that the success rate of the bimodal fusion and intent understanding model is significantly higher than that of the single mode model. This shows that on the one hand, the multi-modal fusion algorithm improves the accuracy of user operations; on the other hand, the intent to understand the model can identify the user's operational behavior and perform voice operation reminders. Thus, the model based on bimodal fusion and intentional understanding works better.

5.3 User satisfaction survey

At the end of the experiment, a questionnaire was provided to compare our platform to the user satisfaction of the NOBOOK platform [8]. NOBOOK experimental system shown in Figure 9. The testers were 20 middle school students, and they were allowed to use the two platforms to conduct experiments and issue questionnaires to them. The experimental scene is shown in Figure 10. The evaluation indicators of the questionnaire were experimental operation sense, interface realism, experimental safety sense and learning effect, and the evaluation questions totaled 8. The problem is evaluated by four evaluation criteria: excellent, good, medium and poor [10]. The statistical results are shown in Table 2. The results show that our platform is superior to NOBOOK in terms of operation sense, interface realism, experimental safety and learning effects. Among them, the experimental operation index results have the largest difference, the difference between “excellent” is 20%, which indicates that the multi-modal fusion algorithm based on decision-making layer fusion reduces the user's operating load and has a better operational sense. Secondly, the results of the learning effect indicators are also quite different, with a gap of 15%, indicating that our virtual experimental device greatly reduces the user's memory load and can effectively help students memorize the chemical experiment steps.

Note: Remember that our platform is A and NOBOOK is B.

6. CONCLUSION

The sodium and water virtual chemistry experimental device proposed in this paper can effectively make up for the shortcomings of traditional experimental teaching, such as the difficulty of traditional sodium and water experiments, the lack of experimental items, and the limitations of the teaching environment. The multi-modal fusion algorithm based on decision-making layer fusion solves the problem of perception of user's operation intention in virtual chemistry experiment, and improves the immersion and accuracy of user operation. Compared with the NOBOOK experimental system, our platform greatly reduces the user's operating load and memory load, has higher user satisfaction, and obtains better experimental teaching results and user evaluation.

The proposed virtual chemistry experimental device lays a good foundation for further exploration and in-depth innovative teaching mode, and helps to promote the establishment of virtual chemical simulation experimental platform. In the future, the
virtual chemical simulation experimental platform will become an effective means to become a reform of chemical experiment teaching.

7. ACKNOWLEDGMENTS
This paper is supported by the National Key R&D Program of China (No. 2018YFB1004901), National Natural Science Foundation of China (No. 61603151) and Natural Science Foundation of Shandong Province (ZR2018PF012).

8. REFERENCES
[1] Liu, D. J., Liu X. L., Zhang W., Lu A. F., Huang R. H. 2016. The Potential, Progress and Challenge of Virtual Reality Technology Education Application. Open Education Res. 4 (Apr. 2016), 25-31.
[2] Xu, G. Y. 2018. The Exploration of the Application of Virtual Simulation in Physics Experiment Teaching. In 8th International Conference on Education and Management. Atlantis Press, Paris, Finance. DOI= http://creativecommons.org/licenses/by-nc/4.0/.
[3] Kawulich, B. B., Adriana, D. A. 2019. Teaching qualitative research methods with Second Life, a 3-dimensional online virtual environment. Virtual Reality. Res. 12 (Dec. 2019), 375–384.
[4] Peixoto, B., Pinto, D., Krassmann, A., Melo, M., Cabral, L., Bessa, M. 2019. Using Virtual Reality Tools for Teaching Foreign Languages. In WorldCIST'19 2019 Conference: Advances in Intelligent Systems and Computing.
[5] Xu, D., Tao, S., Liu, Y.S. 2012. Research and practice of electronic medical records based on voice cloud. China Digital Medicine. Res. 3 (May. 2012), 15-18.
[6] Lin, F. T., Luo, Y., Kong, F. L., Zhang, Y., Qiao, Y. T., Qian, Y. H. 2018. A design of intelligent robot voice interaction system based on cloud platform. Electronic test. Res. 1 (Jan. 2018), 40-42.
[7] Dahl, G. E., Yu, D., Deng, L., et al. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans on Audio Speech & Language Processing. Res. 1 (Jan. 2012), 30-42.
[8] Xie, B. Y. Application of NOBOOK Virtual Laboratory in Primary Science Class. 2017. Experimental Teaching and Instruments.
[9] NOBOOK System. https://www.nobook.com/huaxue.html.
[10] Hu, Z. H., Wang, Y., Wang, C. G., Li P. 2019. Construction and thinking of civil engineering virtual simulation experiment teaching center. Experimental Technology and Management. 5 (Oct. 2019), 218-220. DOI= https://doi.org/10.16791/j.cnki.sjg.2019.10.053.