Multimodal interaction design and application in augmented reality for chemical experiment

Mengting XIAO\textsuperscript{1,2}, Zhiquan FENG\textsuperscript{1,2*}, Xiaohui YANG\textsuperscript{1,2}, Tao XU\textsuperscript{1,2}, Qingbei GUO\textsuperscript{1,2}

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

*Corresponding author, ise_fengzq@ujn.edu.cn

Received: 27 March 2020          Accepted: 20 June 2020

Supported by the National Key R&D Program of China (2018YFB1004901); the Independent Innovation Team Project of Jinan City (2019GXRC013).

Citation: Mengting XIAO, Zhiquan FENG, Xiaohui YANG, Tao XU, Qingbei GUO. Multimodal interaction design and application in augmented reality for chemical experiment. Virtual Reality & Intelligent Hardware, 2020, 2(4): 291—304
DOI: 10.1016/j.vrih.2020.07.005

Abstract  Background  Augmented reality classrooms have become an interesting research topic in the field of education, but there are some limitations. Firstly, most researchers use cards to operate experiments, and a large number of cards cause difficulty and inconvenience for users. Secondly, most users conduct experiments only in the visual modal, and such single-modal interaction greatly reduces the users’ real sense of interaction. In order to solve these problems, we propose the Multimodal Interaction Algorithm based on Augmented Reality (ARGEV), which is based on visual and tactile feedback in Augmented Reality. In addition, we design a Virtual and Real Fusion Interactive Tool Suite (VRFITS) with gesture recognition and intelligent equipment. Methods  The ARGVE method fuses gesture, intelligent equipment, and virtual models. We use a gesture recognition model trained by a convolutional neural network to recognize the gestures in AR, and to trigger a vibration feedback after a recognizing a five-finger grasp gesture. We establish a coordinate mapping relationship between real hands and the virtual model to achieve the fusion of gestures and the virtual model. Results  The average accuracy rate of gesture recognition was 99.04%. We verify and apply VRFITS in the Augmented Reality Chemistry Lab (ARCL), and the overall operation load of ARCL is thus reduced by 29.42%, in comparison to traditional simulation virtual experiments. Conclusions  We achieve real-time fusion of the gesture, virtual model, and intelligent equipment in ARCL. Compared with the NOBOOK virtual simulation experiment, ARCL improves the users’ real sense of operation and interaction efficiency.

Keywords  Augmented reality; Gesture recognition; Intelligent equipment; Multimodal Interaction; Augmented Reality Chemistry Lab

1  Introduction

Virtual experiments are important in field of information intelligence\textsuperscript{10}. It is also an important research area in Human-Computer Interaction (HCI). Virtual teaching methods adopt Augmented Reality (AR)
simulations that realize transformations from two-dimensional (2D) space to three-dimensional (3D) space. Such methods are conducive in improving students' interests, and they also enrich the learning experience. AR technologies include Simultaneous Localization and Mapping (SLAM), card mark recognition, and gesture interaction technologies. In the field of education, most AR studies are based on mobile phones, iPads or computers, and do not need any Head Mounted Displays (HMDs). This is conducive to user-friendly operation. Sun used the Vuforia Software Development Kit (SDK) to identify marked cards, and constructed an AR learning environment with sound animation system, gesture interaction, particle effects, real-time color mapping, and game interaction functions. It has been shown in the study that learning in AR stimulates users' imaginations. Fidan et al. developed a marked AR technology based software called FenAR to support learning activities in the classroom. Studies have shown that incorporating AR into learning activities can improve the academic performance of students. However, users operate multiple cards in experiments, which makes the operation is complicated, and interaction is limited to that with the virtual model by means of external objects. Interactivity is ignored, especially in the study and understanding of middle school chemistry experiments, which cannot be learned without actual hands-on operation.

AR methods based on gesture recognition have been proposed to enhance hands-on operation. Dave et al. combined AR and gestures, and used adaptive hand segmentation and pose estimation methods to recognize gestures in virtual experiments, where users could simulate experiments with their own hands. Ibili et al. developed an AR geometry teaching system with gesture operations, and verified the individual differences in 3D thinking skills of middle school students. They pointed out that personalized, dynamic, and intelligent learning environments are particularly important. Rani et al. proposed an interactive AR system that enabled students to naturally manipulate 3D objects directly using gestures. Studies have shown that systems based on gesture and AR interaction are easy to use and attractive. However, these studies focus only on single modal gesture interaction with heavy operation loads, and the interaction efficiencies are also low. For example, if only gestures are used in simulated chemistry experiments, complex and similar gestures might be misrecognized, thus increasing the operation load. This leads to a reduction in the interaction efficiency and operation.

Gesture recognition is an essential part of achieving AR effects. Gesture recognition methods include deep learning based methods, Hidden Markov Models (HMM), and geometric features. These methods are used to implement gestures and information transmission between visions. Such methods have also been used to realize information transfer between gesture and vision. Wu et al. used the Deep Belief Network (DBN) and 3D Convolutional Neural Network (3DCNN) fusion, and provided the gesture classification probability as input into the HMM model to realize gesture recognition. Elmezain et al. used the HMM model, and recognized the dynamic trajectories of gestures. Priyal et al. used matrix feature normalization to identify the geometry of gestures. Liang et al. used the random forest method to classify gestures, and operated gestures on 3D virtual objects to achieve a seamless fusion of real scenes and virtual objects. According to their research, deep learning methods have promising performance in gesture recognition problems due to their strong fit, and are advantageous over other methods. However, the efficiency of gesture recognition is relatively low in practical applications.

In virtual experiments, the virtual effect presented by AR has strong immersion. The fusion of physical and virtual objects and the fusion interaction of multiple modalities can further improve user interactivity. However, the experiments used in previous studies are not easy to operate, and gesture recognition is inaccurate with heavy operation loads. To address these challenges, we present the following contributions of this study:

1. We use gesture recognition instead of card mark recognition in AR, and propose Multimodal
Interaction Algorithm based on Augmented Reality (ARGEV), based on gestures and sensors. Using Microsoft Kinect, ARGEV turns complex AR tasks into simple coordinate transformation problems, achieving the fusion of users’ hands, virtual scenes, and sensors. It solves real-time interaction of gestures and virtual objects, and improves users' interaction efficiency.

(2) We combine gestures and intelligent equipment to design a virtual and real fusion interaction tool suite (VRFITS), which solves the interaction between physical objects and virtual models, and then triggers perceptual feedback when gestures interact with the virtual model, enhancing the user's sense of real operation.

(3) An Augmented Reality Chemistry Lab (ARCL) is designed, which is operated through the interaction of users' hands, intelligent equipment, and a virtual model.

2 Multimodal interaction method based on AR

VRFITS includes intelligent equipment and multimodal interaction. With visual and tactile feedback, the intelligent equipment can enhance the realism of user operations, and the gesture interaction method is used to identify the users' gestures. Then, a vibrational feedback is triggered by the intelligent equipment.

2.1 The framework of VRFITS

The structure of the framework includes the model establishment stage, gesture recognition and interaction stage, and system application phase stage (Figure 1). In the model building stage, we process the gesture depth map, and deep learning is used to train the gesture recognition model through a Convolutional Neural Network (CNN). In the gesture recognition and interaction phase, we use the gesture recognition model to recognize gestures, and input the gesture depth map to the model. Then, we achieve consistent gesture recognition results through coordinate binding in virtual scenes and real scenes. Fusion of the gestures and virtual model triggers vibration in the intelligent equipment. In the system display stage, we implement ARCL.

![Figure 1 Framework of the multimodal interaction method.](image)

2.2 Design and interaction method for intelligent equipment

Intelligent sensing equipment such as somatosensory devices, Google glass, Microsoft HoloLens and Kinect etc. have become abundant in recent times. Users can use such sensing devices for learning or home entertainment. Researchers can also develop new research based on sensing devices, but the price is very high, which is unrealistic for a large number of teaching classrooms in the Midwest of China. In addition, most equipment cannot be reused in real experiments, which results in wastage. Our design addresses this
by using intelligent equipment with inexpensive sensors.

VRFITS uses sensors in the intelligent equipment (Figure 2) to detect changes in external signals. It connects to the I/O port of the STM32103 main control chip through the signal output port, and uses the serial transmission mode to transmit signals to a computer through the EIA-RS232 module. Finally, the serial data are read in the Unity3D platform. In the processing of the vibrator, the serial port sends the data from the Unity3D platform, and the information is transmitted to the STM32103 main control chip through the serial port to control the vibrator.

The intelligent equipment includes an intelligent ring and two touch sensors, and the approximate cost to make it is 30 Yuan. We set up a vibrator on the intelligent ring, and during the experiment, we put the ring on the little thumb finger of the user's right hand. When the user grabs a virtual object, the sensor induces vibration. The specific perceptual processes (in no particular no order) are as follows:

1. If all five fingers of a user grab a virtual object, the gesture is recognized. The system sends an input "00" to the serial port, and the information is transmitted to the STM32103 main control chip through the serial port. The vibrator sensor shock is set for 5 seconds, and no vibration is induced if all five fingers do not grab the virtual object.

2. When the user's hand touches the tactile sensor, the perceived touch intensity of is calculated, and the average touch intensity is obtained through repeated measurements. If a touch intensity is greater than the average value, the system receive signal Noe, if the Noe=1 indicate the buttons of start experiment and the Noe=2 indicate the buttons end experiment. If there no signal received, it means that semantics are not expressed.

The intelligent equipment is suitable for AR and Virtual Reality (VR) experimental scenes with gesture recognition, and combines gesture behavior with the vibrator to trigger tactile feedback.

2.3 Gesture recognition method

2.3.1 Gesture data preprocessing

We take a chemistry experiment as an example of the virtual experiment system, and count the types of natural gestures used by students or teachers in conducting the experiments. We investigate and design six gestures commonly used in the experiment. First, we use Kinect to collect depth maps of the human body, and collect 10000 pieces of six type. Then, we obtain the depth information and the coordinates of the centroid position of the human hand by Kinect to segment the gesture depth map from the collected body image. A point 3cm from the centroid position is used as the threshold value. If the distance is larger than the threshold value, the human hand area will be exceeded. Then, the human hand area is cut to obtain a gesture depth map of $200 \times 200$ pixels. For simplicity, we consider similar gestures such as a two-finger spread and a three-finger stretch as representing the same semantics. The pre-processed six gesture depth maps and their definitions are shown in Table 1.

2.3.2 Gesture recognition method based on CNN

Then, we build the AlexNet network structure model. The AlexNet structure of CNNs is advantageous, as it can learn richer and higher-dimensional image features. AlexNet uses random inactivation dropout and
data enhancement methods to effectively suppress overfitting, and uses the Rectified Learning Units (ReLU) function instead of the sigmoid function as the activation function. Therefore, for the six kinds of static gesture depth maps, we train a gesture recognition model based on AlexNet. The network structure includes five convolutional layers, three pooling layers, three fully connected layers, and a Softmax classification function. We choose each gesture 10000 depth maps, and each gesture depth map is initially set to resolution of $200 \times 200$ pixels. Then, we obtain the gesture recognition model through the AlexNet network. The network structure is shown in Figure 3.

In AlexNet, the optimal number of iterations epoch set 20000, the number of each training or test batch size is 20, the type of padding is SAME, the value dropout is 0.8, and the generalization performance is evaluated every 20 depth maps. The step size and convolution kernel size of each layer is shown in Figure 3. The process of training the gesture recognition model is as follows:

(1) Using Kinect, we obtain the depth information and depth maps of human bone nodes, and evaluate the human hand area according to the threshold value to generate a gesture depth map.

(2) We divide the dataset into a training set and a test set in the ratio is 7:3.

(3) The training set is input into AlexNet, and the gesture depth features are extracted by continuously
updating the weights according to the following equation:

\[ x^n_i = f \left( \sum_{j=1}^{n} x^{n-1}_j \cdot w_{mj}^n + b_{mi}^n \right) \]  

(1)

where \( m \) is the number of current layers, \( n \) is the number of neurons in the previous layer, \( w_{mj}^n \) is the connection weight of a neuron \( j \) in layer \( m \) and a neuron \( i \) in the previous layer, and \( b_{mi}^n \) is the \( i \) feature bias after \( m \) convolution layers.

(4) After calculating the Softmax layer, the vector \( v \) is obtained. Each dimension of the vector \( v \) represents the probability of the prediction type. The prediction probability is given by:

\[ P_l = \frac{e^{v_l}}{\sum_{i} e^{v_i}} \]  

(2)

where \( v_i \) represents the \( i \)-th element of the vector \( v \), and \( P_l \) is the predicted probability of the \( l \)-th element in the vector \( v \). \( i \in (0,6] \).

Finally, we obtain a trained gesture recognition model (AlexNet_gesture), and encapsulate and import it into ARCL.

### 2.4 Multimodal interaction method

We use the Kinect RGB sensor to build a real environment. The Kinect depth camera captures the depth map of the hand. We realize AR by establishing a coordinate calibration of the real space and virtual space in Unity3D. The ARGVE method achieves tactile and visual fusion interaction. While conducting the experiment, the user directly observes the experimental phenomena and scenes in AR through a computer screen, without wearing any HMDs. If the user makes a five-finger grab gesture, the vibrational feedback is triggered. The process is shown in Figure 4.

**Figure 4** Multimodal interaction process in AR.

First, we encapsulate the AlexNet_gesture model and denote right hand gesture as Ges_R, and left hand gesture as Ges_L, which are given by:
The AlexNet_gesture model accuracy and loss curve are shown in Figure 5. This article sets R1−R6 and L1−L6 represent the Number 1−6 in Table 1. Then, we call AlexNet_gesture in ARCL to establish consistent hand coordinates and virtual coordinates, given by:

$$\theta = (k_x, k_y, k_z)$$

where \(\theta\) is the depth 3D coordinate under Kinect. According to the mapping of the coordinates of the hand joints in real space and depth 3D coordinates, the mapping relationship between the joint point coordinates and the virtual scene is determined as:

$$\begin{bmatrix}
k_x \\
k_y \\
k_z
\end{bmatrix} = t \begin{bmatrix}
U_x \\
U_y \\
U_z
\end{bmatrix} + \begin{bmatrix}
d_x \\
d_y \\
d_z
\end{bmatrix}$$

where \((U_x, U_y, U_z)\) is the virtual scene coordinate in Unity3D, \(t\) is the ratio corresponding to the 3D coordinates of the real-world scene and the virtual scene, and \((d_x, d_y, d_z)\) is the intercept value at the virtual scene coordinates.

In the ARGEV algorithm, we input gesture depth maps and sensor signals, perform multimodal interaction, and output vibrational feedback and visual effects. Visual effects include designed animations, particle effects, and dumping effects of virtual beakers in Unity3D. The specific gesture interaction algorithm is as follows:

**Algorithm 1: Multimodal Interaction Algorithm based on AR (ARGEV algorithm)**

**Input:** gesture depth map, sensor signal Noe;
**Output:** vibration feedback, visual effects;

1. Use the Kinect depth camera to obtain the \((n-1)\) frame gesture depth map and input it into the AlexNet_gesture model for gesture recognition;
2. The \(n\) frame gesture depth map is obtained, the joint point coordinates \(S_{n-1}(\theta_{n-1})\) and \(S_n(\theta_n)\) of frame \((n-1)\) and \(n\), respectively, are recorded;
3. if \((\text{Noe} = 1)\) then the experiment begins, and virtual equipment appears in the scene.
4. if \((\text{Noe} = 2)\) then the experiment ends.
5. if \((S_{n-1}(\theta_{n-1}) = S_n(\theta_n))\) then
   while \((\text{Ges}_R \cup \text{Ges}_L \neq \emptyset)\) do
     if \((R1)\) then set the \(P_{s,R}\) as three-dimensional coordinate of the virtual model
     \(S_n(\theta_n) = P_{s,R}\), send the “00” data to the microcontroller to trigger the intelligent ring vibration end if
     if \((L1)\) then the effect of the prompt box of the selected experimental equipment appears on the system interface end if
     if \((R4)\) then the current virtual equipment is dumping end if
     if \((R6)\) then drop the current virtual equipment end if
   end while
   end if
6. if \((S_{n-1}(\theta_{n-1}) \neq S_n(\theta_n))\) then return 1 end if

### 3 Experimental results and analysis

#### 3.1 The AlexNet_gesture model training results

During the training process, this article sets 20 iterations to detect the accuracy and loss value changes, and visualize it through the Tensorboard method. The AlexNet_gesture model accuracy and loss curve are shown in Figure 5.
In Figure 5, the accuracy rate in the training process gradually tends to 1, and the Loss value changes from large values to stable values, and then tends to 0. This proves that the trained model is continuously optimized effectively.

### 3.2 Comparative experiment

We design two sets of comparative experiments. The first set is the comparison of the accuracy of each gesture before and after model optimization, and the second set is the comparison of the training results of AlexNet, GoogleNet and VGG16Net models. We use the pre-processed gesture depth map and 3000 test pictures in the comparative sets. The experimental results are shown in Figure 6.

![Figure 5 Changes in accuracy and Loss values during training.](image)

![Figure 6 Comparison of experimental results before and after model optimization.](image)

It can be seen that the average accuracy of gesture recognition after optimization is 99.04%, which is about 2% higher than that before optimization. The recognition effect for similar gestures 2 and 3 is also better than that before optimization. The accuracy of the gesture recognition model obtained by the AlexNet model training is better than the other two network models before or after optimization, and the optimized model is improved by about 1%-3%.

### 3.3 Effectiveness of the ARGEV algorithm

We verify the effectiveness of the ARGEV algorithm by evaluating whether the coordinates of the gesture and the coordinates of the virtual model are consistent. When the user makes a five-finger grab gesture, we record the three-dimensional coordinates of the hand and the virtual model simultaneously.
In Figure 7, we label the 3D coordinates of the gesture trajectories with different colors. The 3D coordinates of Figures 5a and 5b are identical at the same time period, which proves that the gesture recognition algorithm is effective.

![Comparison of 3D coordinates of the real hand and virtual model](image)

**Figure 7  Comparison of 3D coordinates of the real hand and virtual model.**

### 3.4 Application of VRFITS in ARCL

We build an interactive virtual and real fusion environment in the intelligent equipment, real hands, and virtual models. We use an Intel® Core™ i7-8550 CPU, Kinect 3.0, and intelligent equipment. We set up experimental scenarios in Unity 3D, and use C# as the programming language. Finally, we apply the ARGEV algorithm and intelligent equipment in ARCL.

VRFITS can be used repeatedly, and it can also help students avoid the dangers in conducting experiments unsupervised. We aim to help students focus on conducting chemistry experiments, enhance observation and learning, and to solve problems such as difficulties and fears associated with experiments, and lack of experimental reagents. We validate ARGEV through an example natrium and water reaction chemistry experiment in ARCL. In the natrium and water experiments, it is considered that the user makes the following gestures: grab, putting down, and dumping. Therefore, in order to facilitate operation in ARCL, we choose three gestures for experimental verification. The flowchart of ARCL is shown in Figure 8.

Natrium and water experiments are the main chemistry experiments in middle school chemistry classes. However, the appropriate amount of natrium and water reaction to generate gas during the user experiment, and a large amount of natrium and water reaction to explode, making the experiment difficult to observe and operate. In order to allow students to better experience the experimental process, this paper takes the natrium and water reaction as an example to present the experimental mechanism in VRCL. In the virtual scene, we add prompt windows, virtual experimental equipment, particle effects, and animation effects to enhance the immersiveness during the experiment. The effect of the experimental operation is shown below.

In Figure 9, the red box indicates operation prompts and scene effects, and the yellow box indicates user operation behavior. After the user touches the start key to begin the experiment, the system presents the AR scene. In the prompt box, the user selects the experimental equipment with the five-finger grab gesture, and the intelligent ring vibrates (a) User dumps the virtual breaker (b) then take out the virtual equipment needed for the experiment.
In Figure 10, it is seen that (a) the user uses a virtual knife to cut the natrium block (b) uses tweezers and takes a small piece of natrium and puts it into a beaker filled with water (c) the user can observe five phenomena of natrium and water reaction and add real video verify authenticity (d) the user selects the beaker again, and (e) puts it on the table, and takes a large piece of natrium into the beaker with tweezers, explosion scenes can be seen (f) finally, the user ends the experiment by pressing the end key.

3.5 User evaluation comparison

We choose natrium and water experiments of the NOBOOK virtual experiment platform\(^{(21)}\) and ARCL to compare performance according to user evaluation (Figure 11). The natrium and water experiment of NOBOOK is a VR simulation experiment, and the system uses a mouse as the input device.

We invited ten teachers and thirty students to complete the evaluation for NOBOOK and ARCL. In order to verify that the experimental system conforms to teaching applications, we set the following seven
Mengting XIAO et al: Multimodal interaction design and application in augmented reality for chemical experiment

comparative aspects: "teaching evaluation," "experimental interest," "experimental interaction," "learning effect," "system stability," "experimental hints," and "ease of operation" as VET_P, and set the VET_P to VET_P1–VET_P7 in order. After the operation, ten teachers compared the two systems using VET_P, and each score was divided into 5 levels, by the contrast score was increased in order (Figure 12). The NOBOOK virtual experiment platform as A and the ARCL as B. We use the ANOVA to assess the significance of each factor (Table 2). The significance α is 0.05.

The teachers were of the opinion that the results of both methods were very satisfactory, and expressed that they could correctly operate the experiments and observe the experimental phenomena. However, the teachers' evaluation of the two experimental patterns are very different, the average value of the ARCL is 29.42% higher than NOBOOK experiment. The teachers' evaluation on VET_P5 are different, and it can be seen in Table 2 that F is less than F crit, which shows that the two systems are relatively stable. For VET_P1, VET_P3, and VET_P7, the overall evaluation of ARCL was 28% higher than that for NOBOOK, and it can be seen that the ARCL operation is simpler and more convenient for teachers to teach than the
virtual and real fusion interaction improves the users’ interaction with the virtual model and the overall cognitive operation load of ARCL is reduced by 27.42%. This proves that the VRFITS interaction efficiency is improved.

It can be seen that the two experiments on the evaluation and of MD are different, which means that the user consumes less mental in the operation experiment. However, in the other five indicators, the evaluation score of ARCL is significantly lower than that in NOBOOK, and the overall cognitive operation load of ARCL is reduced by 27.42%. This proves that the VRFITS interaction efficiency is improved. Thus, virtual and real fusion interaction improves the users’ interaction with the virtual model, and the immersion of the operation experiment.

4 Conclusion and discussion

We propose VRFITS, which contains an intelligent equipment and gesture interaction method. The suite is...
suitable for any AR experiment with gesture operation. We achieve the combination of gestures, sensors, and virtual models in AR. An intelligent equipment and a gesture interaction method assist each other, and the gesture can trigger vibrational feedback. In addition, we design and implement a prototype system called ARCL. According to user evaluation, ARCL increases the interactivity and real sense of operation in virtual experiments, compared to NOBOOK, while reducing the user operation load, and improving the user interaction efficiency. In addition, compared with AR card recognition of the Vuforia SDK, ARCL discards multiple card operations, and uses different gesture commands to trigger different virtual models, which makes the operation more convenient and effective.

However, our work has certain limitations. On the one hand, there are relatively few types in gesture recognition, so there is a lack of gesture types in the interactive process of users in virtual experiments. On the other hand, in the virtual chemistry experiment system, the particle effect, animation effect, and virtual model rendering effect in the experimental scene are not prominent, and the interface effect of the system should be improved in the future.

References

1. Collazos C A, Merchan L. Human-computer interaction in Colombia: bridging the gap between education and industry. IT Professional, 2015, 17(1): 5–9  
DOI:10.1109/mitp.2015.8

DOI:10.1109/bigmm.2017.62

DOI:10.1016/j.cmpb.2018.02.006

DOI:10.1109/vr.2019.8798358

DOI:10.1016/j.ijinfomgt.2018.03.004

6. Sun C X. The design and implementation of children's education system based on augmented reality. The Shandong University, 2017

DOI:10.1016/j.compedu.2019.103635

DOI:10.1007/978-981-13-0224-4_35

DOI:10.1080/0020739x.2019.1583382

Research on the application of virtual simulation experiment in physics experiment teaching of senior high school.


DOI:10.1097/sla.0000000000003173