Taylor & Francis
Taylor & Francis Group

International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

MRLab: Virtual-Reality Fusion Smart Laboratory Based on Multimodal Fusion

Hongyue Wang, Zhiquan Feng, Xiaohui Yang, Liran Zhou, Jinglan Tian & Qingbei Guo

To cite this article: Hongyue Wang, Zhiquan Feng, Xiaohui Yang, Liran Zhou, Jinglan Tian & Qingbei Guo (2023): MRLab: Virtual-Reality Fusion Smart Laboratory Based on Multimodal Fusion, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2023.2227823

To link to this article: https://doi.org/10.1080/10447318.2023.2227823

Published online: 04 Jul 2023.

Submit your article to this journal

View related articles

View Crossmark data

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2023.2227823

Taylor & Francis
Taylor & Francis Group

MRLab: Virtual-Reality Fusion Smart Laboratory Based on Multimodal Fusion

Hongyue Wang[a,b], Zhiquan Feng[a,b], Xiaohui Yang[a,b], Liran Zhou[a,b], Jinglan Tian[a,b], and Qingbei Guo[a,b]

[a]School of Information Science and Engineering, University of Jinan, Jinan, China; [b]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan, China

ABSTRACT

During the COVID-19 pandemic, online classes became the only option for many students. The main challenge for these classes was conducting risky and complex chemical or biological experiments in a domestic environment. To address this challenge, a smart experiment system called MRLab was developed. MRLab used wearables such as a smart glove and head-mounted device to record sensory data and a multimodal hybrid fusion model GVVS to interpret the user's experimental intent, which essentially transforms the user's abstract behavioral actions into a probabilistic set of experimental intent that can be computed. Different experiments in MRLab used different libraries of experimental intents. The SrNet model in GVVS was used to estimate the probability of the user's gesture behavior generated from the smart glove, while the SIPA algorithm compared speech information entered during the experiment with the experimental intent library to estimate the probability of the user's intent. At the same time, the scene visual channel monitored the information about the object the user intended to operate, with the SVF algorithm computing the probability of the intended object in real-time. The results from ANOVA and post-hoc comparative testing conducted on 21 volunteers revealed that MRLab outperformed other experiment modes, including WEB, AR, and VR, with a higher intention understanding rate, efficiency, and user satisfaction. Therefore, MRLab proved to be a useful alternative to traditional physics laboratory experiments during the pandemic, along with being an additional teaching tool for remote learning purposes.

KEYWORDS
Smart glove; mixed reality; virtual-reality fusion interaction; multimodal fusion; intent understanding

1. Introduction

Many secondary school experiments have issues such as reagent contamination, hazardous operations, and high raw material consumption. Additionally, students in educationally underdeveloped areas often lack the resources and time for experimental education. During epidemics, experimental teaching becomes even more challenging to conduct online. Therefore, to support the growth of intelligent education, extended reality (XR) technology has become increasingly popular in recent years (Papakostas et al., 2023).

Secondary school students often use extended reality (XR) technology to complete virtual experiments. Three XR technologies are commonly used: augmented reality (AR), virtual reality (VR), and mixed reality (MR). Firstly, AR allows students to interact with 3D models, improving their understanding of concepts through 3D visualization and eliminating their cognitive gaps between the actual world and data spaces (Chen & Liu, 2020). However, AR's experimental environment depends on recognition cards, causing interaction and recognition problems. Secondly, VR provides an immersive and interactive learning experience (Ozdemir & Ozturk, 2022), but manipulable objects in VR experiments are virtual, lessening the perception of actual operation. Meanwhile, MR provides a brand-new visual environment where real and virtual objects can interact in the same space, pushing development and generating new ideas in various fields, including education (Luo et al., 2020), medicine (Silva Jennifer et al., 2018), and entertainment (Hammady et al., 2020). Therefore, MRLab, a new interactive environment that uses MR technology, satisfies the demands of immersive, real-world experience and natural interaction for experimental instruction.

In virtual and reality fusion experiments, gesture behavior is a crucial component in determining user intention. However, traditional vision-based gesture recognition is often impacted by complicated and obstructed environments, making it difficult to meet experimental requirements. To overcome this challenge, this study firstly created a multi-sensor fusion smart glove (Figure 1) inspired by the TAGLOVE (Cai et al., 2020), this is the hardware basis for assessing and determining the user's experimental operations intention in MRLab.

The main objective of this study is to assist students in completing virtual-reality fusion experiments using wearables in an MR environment. However, the sensor, speech, and scene visual data generated by the students during the experiment are dispersed in time and isolated in space, making it challenging to translate this abstract data into the experimental intention that MRLab can understand.

CONTACT Zhiquan Feng ✉ ise_fengzq@ujn.edu.cn School of Information Science and Engineering, University of Jinan, Jinan, China

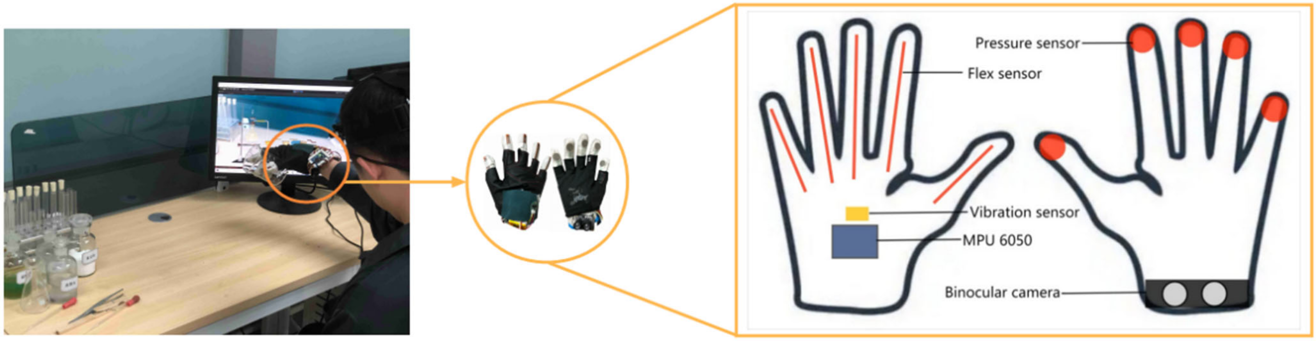

Figure 1. (Left) An image showing a user operating chemical experiments using a smart glove and head-mounted device; (Middle) A close-up of the front and back of the smart glove; (Right) A schematic showing the types and locations of sensors incorporated in the smart glove.

Consequently, the following novel contributions were proposed by this paper:

1. A smart glove prototype was developed to collect the user's gestural motion signals and scene visual data in an MR environment. This smart glove (available at approximately $100) contains a cheap commercial sensor and a binocular camera, which allows for precise motion and visual detection.
2. A multimodal hybrid fusion model GVVS, incorporating a sensor channel, speech channel, and scene visual channel, is established to quickly and accurately categorize user intentions during user experimental operations. This model enables the collection and processing of continuous or discrete data information produced during the experiments and simultaneously processes data from multiple sources using feature layer fusion and decision layer fusion, make it simpler to improve the accuracy and speed of categorizing user intentions in MRLab.
3. A head-mounted device is used to establish a virtual-reality interaction channel and capture the user's head posture. Meanwhile, a display shows the smart experimental scene in MR environments, allowing users to operate experimental instruments and chemicals in the real world while observing corresponding experimental phenomena in the virtual world. This feature satisfies users' desire for immersion and real-world operation.

The paper is structured as follows: Chapter 2 provides an extensive explanation of related work. In Chapter 3, the multimodal fusion model and smart glove prototype design are described. Chapter 4 focuses on analyzing and discussing system data and experimental findings. The discussion, future work, and conclusions are presented in Chapters 5 and 6.

2. Related work

2.1. Experimental teaching of virtual simulation

Virtual simulation technology was gradually integrated into experimental education processes toward the end of the 20th century. This technology offers the advantage of recreating real-world objects and replicating difficult-to-observe events when compared to traditional hands-on experimental teaching (De Jong & Van Joolingen, 1998). However, during the early days of integrating virtual simulation technology in education, several challenges regarding expensive and massive equipment, as well as a high demand for professional and technical employees, coupled with insufficient attention given to the innovation of experimental content were encountered. Recently, there have been significant advances in computer graphics and hardware, and as a result, the focus of research has shifted from obtaining superior digital equipment to the production of experimental content employing advanced technology.

The web-based virtual experiment platform has significantly reduced the cost of virtual simulation experiments, allowing students to conduct frequency modulation experiments regarding communication principles through a web browser (Chi Chung et al., 2001). However, using the conventional Windows, Icons, Menus, and Pointers (WIMP) interface for computer interaction in experimental education may lead students to lose sight of real-world operation experience (Beaudouin-Lafon, 2000; Jacob et al., 2008). Fortunately, the advent of the internet and the proliferation of mobile devices offer effective solutions to these issues. Virtual modeling can be accomplished, and interactive applications of actual physical objects can be developed using smartphones and Augmented Reality (AR) technology (Heun et al., 2013). This approach offers the advantage of using mobile devices to enhance students' interaction experience with the physical world, improve their learning efficiency, and leverage their familiarity with the way media devices are used for interaction (Han et al., 2019). However, this method has limited tactile feedback, does not effectively handle the relationship between real-world and virtual object occlusion, and may make the interaction process unnatural since users have to rely on mobile devices frequently during operation.

With the development of virtual simulation technology, "real-virtual continuum" was introduced as a concept that emphasized MR's focus on the interaction and integration of the virtual and real worlds, while AR augments the real world, and VR enhances the virtual world (Wang et al., 2020). On the basis of MR technology, the application of tangible user interface (TUI) is gradually developing

(Ishii & Ullmer, 1997). In the TUI environment, physical objects can be fully matched with virtual superimposed information, satisfying natural user interaction (Azmandian et al., 2016). However, TUI's interaction technology is prone to registration tracking errors and significant interaction delays, casting doubt on its reliability. Currently, the 3D interactive system created with the Kinect device addresses these issues and allows direct interaction with virtual objects in the 3D user interface (Hilliges et al., 2012). This interactive method lets users interact with computers in both physical and virtual 3D spaces, integrating virtual and actual teaching environments (Lee et al., 2013).

Virtual simulation experiments can help users comprehend complex information and increase their interest in learning while minimizing risks. Comparatively, MRLab focuses on MR technology and 3D user interface (3DUI) interactive methods, utilizing a head-mounted device to create an interface among the virtual world, the real-world, and users. This device also enhances the sense of realism in the user experience by creating an interactive feedback path between the user, the real-world, and the virtual world.

2.2. Multimodal fusion

The majority of the experimental environments previously discussed are built around single-channel interaction, which is unsuitable for experimental instruction that involves a lot of hands-on work. As a consequence, in MR environments, multi-channel interaction has been developed to support users in interacting with computers through various channels, such as voice, gestures, visuals, sounds, and smells (Mistry et al., 2009). This approach achieves the addition and complementation of interaction information between different channels, aiding learners in obtaining a natural 3D interaction experience during experiments (Wanick et al., 2018; Wei et al., 2011). The challenge for this research approach is to address the efficiency, accuracy of user's intentions, and naturalness of the interaction, and processing user input data. Various studies have shown that when compared to relying on single modes of information, fusing multi-channel information can efficiently and reliably extract the intention behind user operations, increase the naturalness and efficiency of the interaction (Alzubi et al., 2023; Sun et al., 2021).

There are four main multimodal fusion techniques: data layer fusion, feature layer fusion, decision layer fusion, and hybrid fusion. Data layer fusion, which is the lowest-level fusion method, involves processing and filtering raw data to facilitate handling, storage, and transmission. Techniques like data noise reduction and data compression are used to optimize the data. SVM models (Zhao et al., 2007) and deep learning methods (Zhang et al., 2018) can be used for the fusion of multi-source heterogeneous data to improve data accuracy. Nonetheless, only using data layer fusion challenges the determination of feature correlations across incoming input, thereby making it difficult to determine the purpose of user actions. With the advancement of deep learning, feature layer fusion has significantly increased the recognition rate in fields like target detection (Bai et al., 2022) and picture recognition (Sun et al., 2005). Feature layer fusion extracts specific features from various modalities and transfers them using a transformation to a high-dimensional feature vector, enabling the training model to classify the user's intended behavior. The main benefit of this method is increased accuracy in identifying intent by analyzing relevant features of various model information, but identifying the most significant features takes time. Decision layer fusion fuses the outcomes of several modalities to increase decision accuracy by utilizing different learning algorithms for each model information (Chen et al., 2004). However, it cannot effectively learn the impact of feature relationships between modalities on intent recognition. Hybrid fusion uses both feature and decision-level fusion methods to obtain the user's intention (Zhou et al., 2020).

Considering that the sensor, voice, and visual information in MRLab are spatiotemporally isolated, this study proposes the GVVS hybrid fusion model for multimodal information. The model includes (1) SrNet feature fusion model for the sensor channel, which produces the hand behavior probability used for user experimental operation, (2) SVF algorithm for scene vision channel, which uses a head-mounted device and a glove's wrist camera to track the probability data of the experimental object that the user wants to acquire in real-time, and (3) SIPA algorithm of the speech channel, which uses the text similarity matching approach to determine the probability of the user's intention. Finally, at the decision-making level, the improved information volume weighting method is utilized to fuse the three-channel probability information to generate the user's operation intention in Mixed Reality Lab.

2.3. Smart wearables

The technology of smart wearables combines sensing, data processing, communication, and interaction (Mann, 1998). It gathers data through user-device interaction and store or transmit it in real-time. Due to its miniaturization, light weight, and wearability features, it is easy for users to use and transport. Smart wearables are typically separated into head-mounted, body-dressed, hand-worn, and foot-worn categories based on the physical characteristics of the human body structure. In the XR environment, various wearable devices are employed depending on the specific application.

Head-mounted wearables, such as glasses and helmets, can be used to assist visually impaired individuals in performing orientation and movement training in virtual reality (VR) environment with the use of smart glasses (Thevin et al., 2020). Additionally, researchers have demonstrated that a virtual character modeling system can replicate the handcrafting process from the physical world in a VR environment using a head-mounted display (HMD) and Leap Motion (Park et al., 2017); Wearables such as coats, underwear, and trousers are embedded with sensors made of conductive fibers and nanomaterials. Researchers have developed a digital clothing prototype, utilizing Sparse Soft Sensors, which can facilitate 3D human body reconstruction

(Chen et al., 2021); In order to enable distant communication between sign and non-sign language users, an AI-based sign language translation glove was proposed to project the results of sign language recognition into the VR environment (Wen et al., 2021); Wearable devices can also be attached to the feet, such as socks and shoes. A smart sock utilizes four soft stretchable sensors based on silk fibroin yarn to perform real-time 3D reconstruction of the foot (Zhang et al., 2020). Compared to traditional devices like computers and mobile devices, wearable technology offers a greater range of human-computer interaction methods, which enhances the user experience and improves the quality of life.

Smart gloves equipped with multiple sensors have shown potential in various applications, such as gesture capture, robot manipulation (Roy et al., 2015), rehabilitation training (Ma et al., 2016), and sign language recognition (Luzhnica et al., 2016). Hand movements also play a critical role in operating the experiment. On the basis of our earlier research (Wang et al., 2022), this study utilizes wearable technologies, such as smart gloves and head-mounted devices, to capture the user's experimental intentions with a high degree of operational freedom in the MRLab. Meanwhile, collaborative processing of data from multiple channels can aid in carrying out virtual and real fusion experimental exercises in secondary school, supported by interactive modes such as sensing, speech, and visual information.

3. Overview of MRLab

This paper presents the development of MRLab, which assists secondary school students in conducting experiments using a smart glove and a head-mounted device. The ability to understand the student's experimental intentions is crucial to the success of MRLab, as it allows for accurate and efficient support for virtual-reality fusion experiments. The experimental intention in this study refers to the user's operational steps during the experiment, which MRLab identifies using the GVVS model in Section 6, and provide appropriate feedback or correction according to the multimodal output module in Section 7.

Multimodal intent identification involves three stages: information gathering, intent analysis, and intent fusion and extraction. Section 2 discusses information gathering, while Sections 3–5 provide a detailed explanation of intent analysis. Section 6 covers the process of intent fusion and extraction.

3.1. GVVS: multimodal hybrid fusion model

The sensor, voice, and scene visual information produced by users during smart experiments in MRLab are often isolated in space and dispersed in time. To address this issue, this study proposes the GVVS hybrid fusion model for multimodal information. The paper discusses a multi-sensor feature fusion model based on *SrNet* for obtaining user hand action probabilities and an SVF algorithm using the improved YOLOv5 technique to derive the probabilities of objects that the user intends to operate from continuous information channel I_cs, including sensor and scene visual information. Discrete speech information is processed using the SIPA algorithm to derive the set of intention probabilities I_vc under the speech channel (discrete information channel) with Baidu speech recognition technology and Chinese lexical analysis technology. These two sets of probabilities are fused at the decision layer in GVVS to generate the final experimental intention. Figure 2 displays the general structure of the GVVS model.

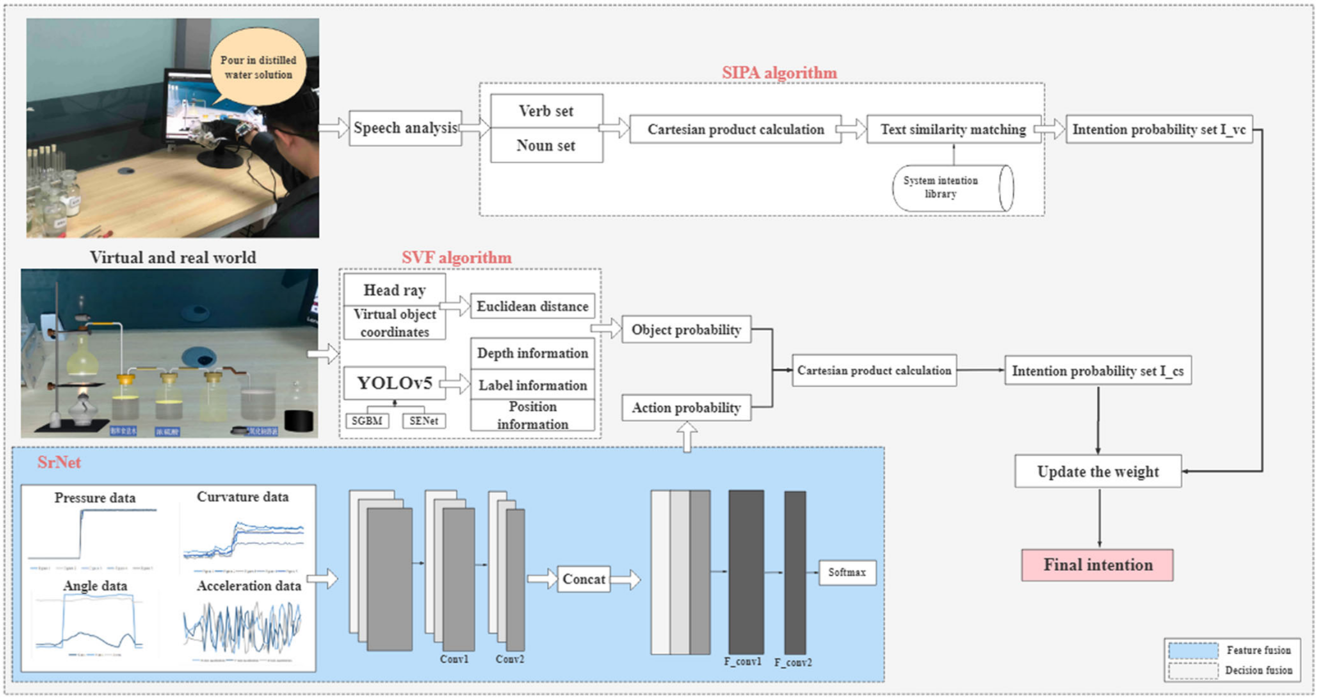

Figure 2. GVVS overall framework diagram.

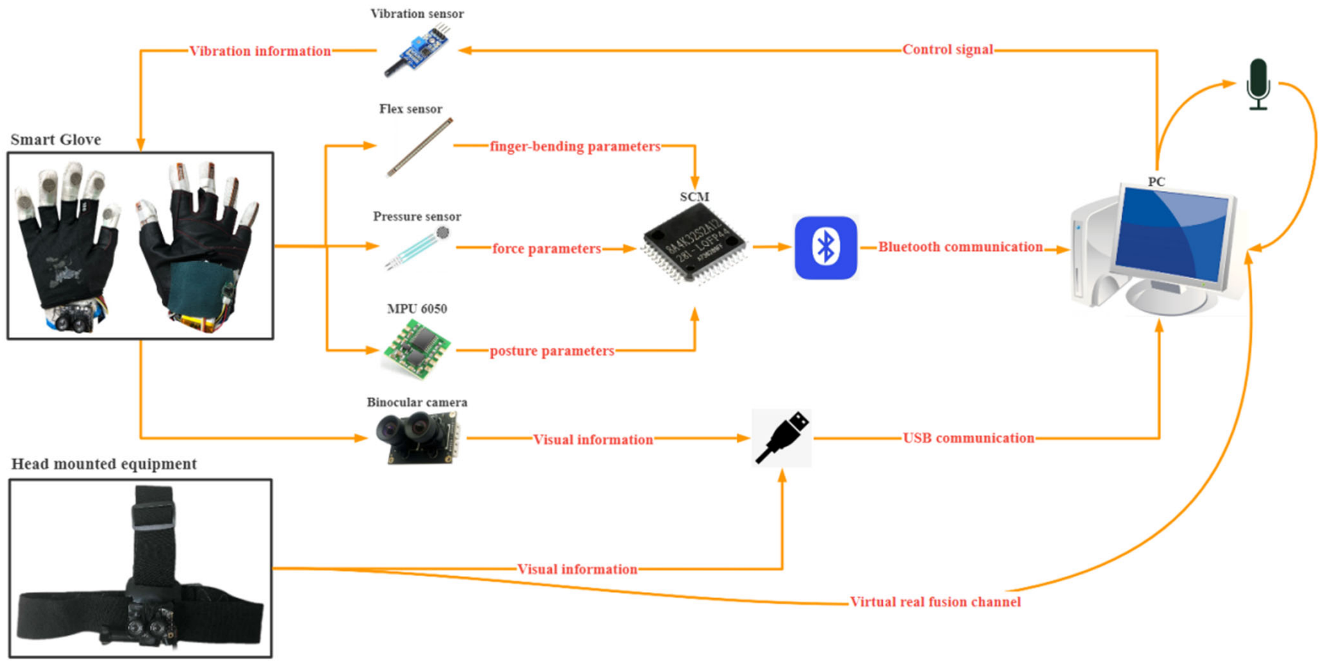

Figure 3. System hardware composition structure diagram.

3.2. Hardware and software implementation

This study aims to design a smart glove with multi-sensor fusion specifically for use in secondary school experiments. The glove is equipped with a microcontroller, sensor set, and binocular camera to capture multi-sensor signals and scene information from the user. The collected data is then used to analyze the user's gesture movements and determine the experimental objects they want to operate.

1. The microcontroller is the central control module used for handling the sensor signals.
2. The sensor set comprises of a vibration sensor, flex sensor, pressure sensor, and posture sensor (MPU 6050).

    - Flex sensor located in the finger section measures the degree of finger bending and maps the changes to the virtual hand in the Unity scene.
    - The MPU 6050, situated on the back of the hand, measures the hand's three-axis angle, angular velocity, and acceleration.
    - A pressure sensor in the finger belly measures the condition of the finger while operating an experiment.
    - A vibration sensor located in the back of the hand provides vibration feedback during the contact process, thereby creating a more genuine user experience.

3. The smart glove has a binocular camera fixed in its wrist portion. This feature allows for real-time information gathering about the experimental scene, addressing the issues commonly found in traditional virtual experiment practices that involve cameras and Kinect devices resulting in obscuration and poor long-distance object recognition accuracy. The smart glove's sensor group and binocular camera transfer sensing and visual data, respectively, to the computer via Bluetooth and USB. Figure 3 illustrates the hardware structure of this system.

MRLab, in contrast to other VR devices such as VR glasses, can achieve 3D registration of digital objects with physical spatial locations using a simple head-mounted device. It enables the interaction of real experimental instruments with virtual objects, allowing users to perform a variety of experiments in smart labs while utilizing technologies like virtual-reality overlay and human-computer interaction. The FPS of MRLab remains stable at 23 frames per second, and it is primarily developed using the Unity engine and C# language.

3.3. SrNet: a multi-sensor feature fusion model based on smart glove

While user behavior recognition employing multiple heterogeneous sensors is widely used in various fields, identifying complex human activities remains a challenge. Human activities typically occur in complex environments where multiple behaviors are carried out simultaneously. Multiple heterogeneous sensors fused for data fusion can also lead to compatibility problems, resulting in low recognition accuracy for concurrent complex activities. This study is limited to the application of user behavior in smart labs, where sensors are used to detect users' experimental actions. Additionally, this paper proposes a feature fusion-based multi-sensor model called *SrNet*, which uses convolutional neural networks to automatically extract features from each sensor's raw data to improve the generalization and robustness of detecting users' experimental operating behaviors.

This model is split into two main sections and employs the thinking of feature layer fusion:

(1) Each of the three sensors ($S_i(0 < i < 3)$) has a corresponding 2-layer convolutional subnetwork. The convolutional subnetwork takes a segment of the sensor time series matrix $V(S_i)$, obtained from a sliding window of t seconds, as an input. Each layer of the convolutional subnetwork undergoes a 2D convolutional kernel with a size of (1,3) to learn the data features. The feature maps $conv_{i1}$, $conv_{i2}$ are then obtained sequentially using Equation (1).

$$conv_{ij} = a_{ij-1}W_{ij} + b_{ij} \tag{1}$$

where a_{ij-1} denotes the output value of the i-th sensor at layer $j-1$, the convolution kernel weight is W, and the bias is b. And then, the $conv_{i2}$ of the depth feature maps of these three sensors are merged to obtain a large depth feature map.

$$X_0 = concat(conv_{02}, conv_{12}, conv_{22}) \tag{2}$$

(2) Next, the large feature map passes through two convolutional network layers to determine correlations between multiple sensor features and obtain F_conv_1 and F_conv_2 in turn. Finally, a *Softmax* classifier is employed in the fully connected layer to classify user actions and output probabilities as indicated in Equation (3).

$$P(action_i) = softmax(F\_conv_2) \tag{3}$$

In this case, this model uses the labels from the actual classification to optimize parameters using a back-propagation algorithm. Additionally, each convolution layer performs batch normalization to avoid vanishing gradients and is accompanied by a non-linear RELU activation function.

3.4. SVF algorithm: decision level fusion algorithm of scene vision channel

Under the scene visual channel, the binocular camera on the smart glove perceives the entire experimental scene. This paper incorporates the SENet attention mechanism to improve perception accuracy and the stereo-matching algorithm SGBM to extract depth information, and finally generates the probability set G of experimental instruments that the user wants to operate by identifying the depth information and category probability of the object Bounding Box, as shown in Figure 4.

Furthermore, the head-mounted device can record both the data related to the user's head posture and serve as a channel for virtual-real-world interaction. Consequently, utilizing the spatial location of the head-mounted device as the coordinate origin, this study proposes a head ray R based on the head-mounted device and YOLOv5 technology to dynamically generate the probability set H of experimental things that the user desires to operate.

The decision-level fusion process in this academic paper utilizes the Bagging concept of integrated learning. Its objective is to fuse two probability sets and output the probability information of the object that the user intends to manipulate through the visual scene channel. The specific algorithm steps are as follows.

Algorithm 1: Scene visual fusion algorithm (referred to as SVF)

```
Input: head ray $R$, set of experimental objects $OBJ$;
Output: set of experimental objects probability under scene visual channel;
1. $While\ OBJ! = Empty\ do$
2. The improved YOLOv5 model is used to obtain the depth information $dep_i$ and the category probability $P(obj_i)$ of object $obj_i$ in the object set $OBJ$,
       $[dep_i, P(obj_i)] = get(obj_i)$
3. Calculate the weights corresponding to $obj_i$ under the visual channel according to the incremental change of the object's depth,
       $W(i) = \frac{\Delta dep_i}{\sum_{m=0}^{n} \Delta dep_m}$
4. Calculate the probability of the object $obj_i$ in the set of experimental objects $OBJ$ under the scene vision channel of the smart glove,
       $P_G(obj_i) = W(i)P(obj_i)$
5. $While$ R! $= Empty\ do$
6. The intersection point $(X, Y)$ is formed between the ray $R$ and the plane where the experimental object is located, and the distance between the Bounding Box coordinate points $(x_i, y_i)$ of all experimental objects and the intersection point is calculated,
       $distance_i = \sqrt{(x_i - X)^2 + (y_i - Y)^2}$
7. Calculate the probability of the object $obj_i$ in the set of experimental objects $OBJ$ under the scene vision channel of the head-mounted device,
       $P_H(obj_i) = 1 - \frac{distance_i}{\sum_{m=0}^{n} distance_m}$
8. Probabilistic fusion of two experimental object probability sets,
       $P(obj_i) = \frac{(P_G(obj_i) + P_H(obj_i))}{2}$
9. End
```

3.5. SIPA algorithm: Speech channel intention probability acquisition algorithm

The users can input speech data to the speech channel any time in the MRLab. However, current research requires users to fix the speech data, such as voiceprint (Nidhyananthan, 2018) or keyword recognition, which can increase the user's psychological burden and conflict with

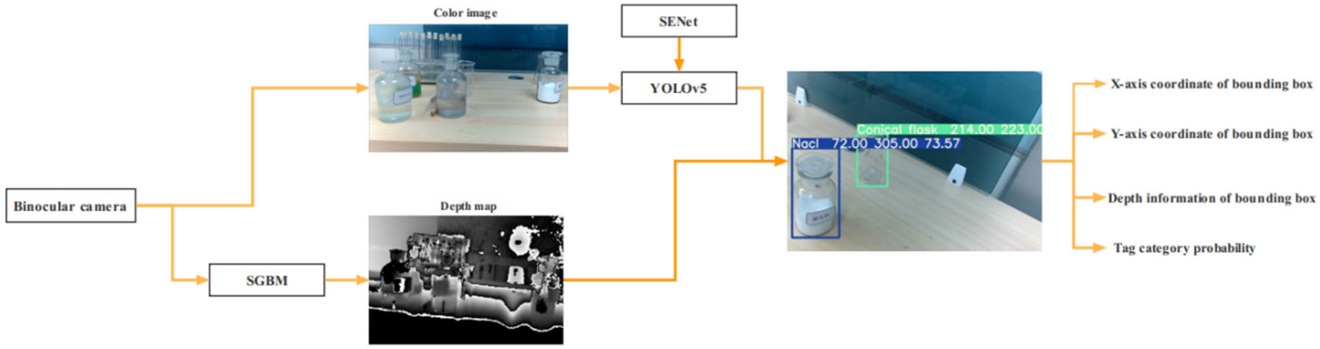
Color image
SENet
YOLOv5
Binocular camera
Depth map
SGBM
Conical flask 214.00 223.00
Nacl 72.00 305.00 73.57
X-axis coordinate of bounding box
Y-axis coordinate of bounding box
Depth information of bounding box
Tag category probability

Figure 4. Scene visual channel.

human-computer interaction principles (Kaptelinin, 1996). Therefore, this paper proposes a probabilistic algorithm to acquire the user's speech intention based on text similarity matching, which allows users to input arbitrary information in a more relaxed manner and enables the system to understand the user's speech intention accurately through probabilistic calculations.

This algorithm builds upon the Baidu Speech Recognition API and Chinese lexical analysis technology. Meanwhile, the user's voice is continuously monitored by MRLab and separated into a set of verbs $\mathrm{s_v}$ and a set of nouns $\mathrm{s_n}$. Furthermore, the set of temporary intention V is obtained through the Cartesian product between $\mathrm{s_v}$ and $\mathrm{s_n}$. Then, the user's experimental intention probability set in the speech channel is determined by matching the text similarity with the experimental intention database *ES*, and the specific algorithm steps are as follows.

Algorithm 2: Speech channel intent probability acquisition algorithm based on text similarity matching (referred to as SIPA)

```
Input: user input speech s, experimental intent library $ES$;
Output: the set of experimental intention probabilities $I\_vc$ under the speech channel;
1. While s! = Empty do
2. Divide the user input speech message s into verb set $\mathrm{s_v}$ and noun set $\mathrm{s_n}$,
       $\mathrm{s_n}, \mathrm{s_v} = \mathrm{Participle(speech)}$
3. perform a Cartesian product operation on $\mathrm{s_n}$ and $\mathrm{s_v}$, and generate a temporary set of intentions $V$,
       $V = \mathrm{Dikaer}(\mathrm{s_n}, \mathrm{s_v})$
4. Matching the instructions $V$ with the experimental library ES for text similarity and updating the set of intention probabilities under the speech channel,
       $I\_vc = \mathrm{P(Matching(V, ES))}$
5. End
```

3.6. Multimodal fusion

Multimodal intent fusion and extraction is necessary for human-computer collaboration and interaction in MR smart experiments. The entire input information from all channels must be integrated by multimodal fusion before the system can authenticate users' intent completely. Therefore, the probability sets of intentions from all three channels are combined at the decision level using the information quantity weights in this research.

Suppose the user is performing in experiments, and the experimental intent library *ES* has *R* experimental intentions, indicated as $I_{1\ldots R} = \{I_1, \ldots, I_R\}$. The system analyses the sensor channel data from the *SrNet* model to determine the likelihood of user action, which is indicated as:

$$P(action) = \{P(\mathrm{Grasp}), P(\mathrm{Release}), P(\mathrm{Pour}), P(\mathrm{Pinch})\}$$

The SVF algorithm analyses the information from the scene visual channel to determine the probability of the experimental objects the user intends to manipulate. Suppose there are *S* experimental objects, the probability is represented as:

$$P(object) = \{P(object_1), P(object_2), \ldots, P(object_S)\}$$

In order to obtain the set of intention probabilities $I\_cs_{1\ldots R} = \{I_{cs1}, \ldots, I\_cs_R\}$ under the continuous information channel, this paper operates $P(action)$ and $P(object)$ as follows:

$$I\_cs = \{P(action) \times P(object) | (action + object) \in ES\}$$

where $(action + object) \in ES$ indicates that the intent resulting from splicing the action and object should be in the experimental intent library of the current experiment. And if this condition is satisfied, the Cartesian product operation can be performed on $P(action)$ and $P(object)$, and finally, the set of experimental intent probabilities I_cs under the continuous information channel is obtained.

If the user does not input any speech information during the experiment, a speech (discrete information) channel intention probability set I_vc is not generated, denoted as $I\_vc_{1\ldots R} = \{I\_vc_1 \ldots I\_vc_R\}$. In this case, we consider the final user intention as:

$$I = \max(I\_cs) \tag{4}$$

If voice information is input, the system requires intent fusion of I_cs and I_vc, which is given by:

$$I = \max(\omega_{I\_cs} I\_cs + \omega_{I\_vc} I\_vc) \tag{5}$$

Each channel varies in importance among them, therefore in this research, we introduce the parameter ω and use the

Table 1. Multimodal output module.

Multimodal output module	Basic needs	Sensory needs	Voice prompt and correction
			Experimental phenomenon display
			Real tactile feeling
			Vibration sensation
		System needs	Smart glove
			Head-mounted device
			Real experimental instrument
			Smart experiment design
	Security and comfort needs		Experimental safety
			Physical comfort needs
			Psychological comfort
	Belongingness needs		Listening and speaking interaction needs
			Experimental Q&A interaction needs
	Esteem needs		Encouragement
			Systematic respect
			Game Entry Mode
	Self-realization needs		Experimental principle learning
			Grade scoring needs

information quantity weight method to dynamically calculate the weight of each channel's intention probability set. However, before calculating the weights, we must determine the coefficient of variation D of these two channels. In conventional practice, calculating the coefficient of variation is achieved by finding the variance and mean of probability sets with R experimental intentions (Bedeian & Mossholder, 2000). The larger the difference between the probabilities in the intention probability set, the higher the accuracy of the channel's intention recognition is demonstrated to be. As a result, it requires a higher weighting. This research improve the classic coefficient of variation solution as follows in order to make the coefficient more compatible with the technique of calculating the experimental intent and to prevent the set's intent probabilities from being too high or too low to produce unrealistic values.

$$\begin{cases} D_{I\_cs} = \dfrac{Mean(g(\mathrm{I\_cs}), R)}{Variance(I\_cs, R)} \\ D_{I\_vc} = \dfrac{Mean(g(I\_vc), R)}{Variance(I\_vc, R)} \end{cases} \tag{6}$$

where $g(\cdot)$ represents the Gaussian distribution of the distance between intention probabilities, as shown in the equation.

$$\begin{cases} g(\mathrm{I\_cs}) = e^{-\frac{(\mathrm{I\_cs}_i - I_{\max})^2}{2\sigma^2}} \\ g(\mathrm{I\_vc}) = e^{-\frac{(\mathrm{I\_vc}_i - I_{\max})^2}{2\sigma^2}} \end{cases} \tag{7}$$

where, while $I_{\max}$ is the maximum reliable probability threshold we define, σ^2 is a constant that limits the range of values. In this paper, the normalization operation is performed on the coefficients of variation to obtain the dynamic weight information under continuous and discrete information channels, as shown in the equation.

$$\omega_{VT}, \omega_{VC} = Normalization(D_{VT}, D_{VC}) \tag{8}$$

where ω_{I_cs} and ω_{I_vc} both belong to $(0, 1)$, and $\omega_{I\_cs} + \omega_{I\_vc} = 1$.

3.7. Multimodal output

To provide users with a positive experimental experience and advantageous learning outcomes, this paper uses Maslow's needs theory to direct and correct users' experimental behavior (McLeod, 2007). According to this psychological theory, physiological needs, security needs, belongingness needs, esteem needs, and self-actualization needs describe the range of human needs. To better align with the demands of virtual experiments, an MVE model that builds upon Maslow's needs theory was presented in the context of smart experiments (Pan et al., 2022). They contend that traditional Maslow's Needs Hierarchy theory (MNH) no longer satisfies the practical conditions of smart experimentation. Thus, they expand the scope of physiological needs to involve basic needs (such as sensory and system needs) and security needs to encompass both safety and comfort needs. This paper introduces a Multi-Modal Output Module that motivates learning by creating a comfortable, safe and realistic virtual environment for students to conduct experiments. A framework of the module is shown in Table 1.

The smart glove-based MRLab simulates multi-sensory experiences such as visual, auditory, tactile and vibration sensations that enable efficient knowledge mastery for students' fundamental sensory needs. The system hardware needs utilize a smart glove and a head-mounted device to fully immerse users in the system. Additionally, on the software side, experiments are created in accordance with established experimental laws, and information augmentation techniques simulate the corresponding experimental phenomena.

Real experiments often pose safety risks associated with explosions, flames, corrosion, and toxic gases. Integrating virtual reality technology into smart experiments can meet the safety and comfort needs of the users, enabling them to concentrate on exploring the experiment's principles without undue psychological pressure.

During actual experimental teaching, students' needs for belongingness are enhanced by teacher-student interaction. In virtual experiments, users often feel isolated, which can impede their transition into the learning and experimental states. To address this issue, the system has incorporated modules which provide listening and speaking interaction and Q&A interaction enabling the users to communicate and feel less isolated.

At the same time, students strive to fulfill their esteem and self-actualization needs through curiosity and encouragement, with minimal assistance. To encourage motivation and reduce negative feedback during the experimental process, the system uses incentives such as encouragement, level scoring, game mode, and other interactive methods. These incentives help develop the needs for esteem and self-actualization, fostering the students' motivation to learn.

4. Experimental results and discussion

4.1. System setup and experimental method

The tests were conducted at MRLab utilizing a smart glove, a head-mounted device, and a computer with an Intel(R) Core(TM) i7-10875H CPU and Nvidia RTX 2060 GPU. The MRLab predominantly uses the Unity engine and C# programming language, and the analysis procedure is based on PyTorch deep learning framework using Pycharm. And, speech channel intention is analyzed through the Baidu speech recognition API and Chinese Jieba word segmentation lexical analysis for recognizing user speech.

Ten volunteers with an age mean of 26.2 and a standard deviation of 4.24, including 6 men and 4 women, participated in the experiment for *SrNet*. They used smart gloves developed in this study to grip, release, pour, and pinch while continuously recording the signal. So, the experiment produced 6382 training and 2735 test datasets saved in *.csv* format, where some data was vulnerable to noise and outliers. Because of the curvature sensor and pressure sensor can only produce five valid data for the five fingers of the glove, but the posture sensor generates six valid data (three-axis angle and acceleration), so other data must be filled with the operation.

The experiments involving the use of the smart glove were approved by the IRB, and all participants in these experiments have provided written informed consent.

4.2. Experimental setup

To compare the learning effects of MRLab with the traditional WEB (Fang et al., 2020), AR/VR (Han et al., 2020), and traditional MR experimental environments (Zeng et al., 2020), this research set up the ablation experimental environments used for comparison, as shown in Table 2.

This paper exemplifies four experiments, "chlorine gas preparation," "charcoal reduction of iron oxide," "ammonia production," and "red phosphorus combustion." As shown in Table 3, volunteers were requested to carry out the major steps of the experiments and record the associated experimental data. From secondary schools, a total of 21 volunteers were recruited, consisting of 10 males and 11 females, without prior experience in VR experiments. The protocol entailed providing an extensive introduction to the VR environment before the experiment began. Each volunteer practiced the experimental scenario for 5 minutes to familiarize them with the interface's capabilities. Figure 5 illustrates each experiment's VR display interface.

Table 2. Four different experimental environments.

Experiment environment	Description
EV1	WEB-side experimental environment (NOBOOK)
EV2	AR/VR experiment environment (VR glasses)
EV3	Traditional MR experiment environment (Kinect/Camera)
EV4	MRLab

4.3. Comparison of algorithm effectiveness

The multimodal hybrid fusion model proposed in this paper originates from a previous study that introduced an MFA multimodal fusion method based on decision layer fusion (Wang et al., 2022). The study conducts a series of comparative experiments where volunteers complete the tasks twice, once in the MRLab environment and once in the environment used in the previous study. Then, the experimental data of completion rate, average completion time, and user satisfaction (measured on a scale of 1–10 points) are counted, as depicted in Figure 6, to explore the correlation between different environments and performance metrics.

As demonstrated in Figure 6, both of the multimodal fusion algorithms exhibit high intent recognition accuracy, with the recognition rates consistently above 94%. Nonetheless, the GVVS model outperforms the MFA algorithm in both intent recognition accuracy and experiment completion rate. The GVVS model improves the intent analysis and fusion process of multiple channels, including speech, sensor, and visual, and also integrates head pose information from the head-mounted device.

Previous studies often rely on devices like standing cameras or Kinect to observe user actions, which may cause potential issues of obstructing visual information and missing small experimental objects. Whereas, the development of MRLab alleviates the physical and mental strain on users, enhances the level of immersion during experiments, and facilitates secondary school students to carry out experiments with high accuracy and efficiency.

4.4. Time performance evaluation

Figure 7 displays the mean times required to complete various key experimental steps in each experimental setting, with the error bars showing the standard error of the mean.

Table 3. Key steps of the four experiments.

Task	Description
1	Add potassium permanganate and dilute hydrochloric acid to the reaction vessel for chlorine gas preparation and collect the chlorine gas produced.
2	Add charcoal powder and iron oxide powder to the reaction test tube, heat the test tube and observe the experimental phenomenon.
3	Heat concentrated ammonia to prepare ammonia gas and test the acidity and alkalinity with red litmus paper.
4	Put the burning red phosphorus into the beaker with oxygen and observe the experimental phenomena.

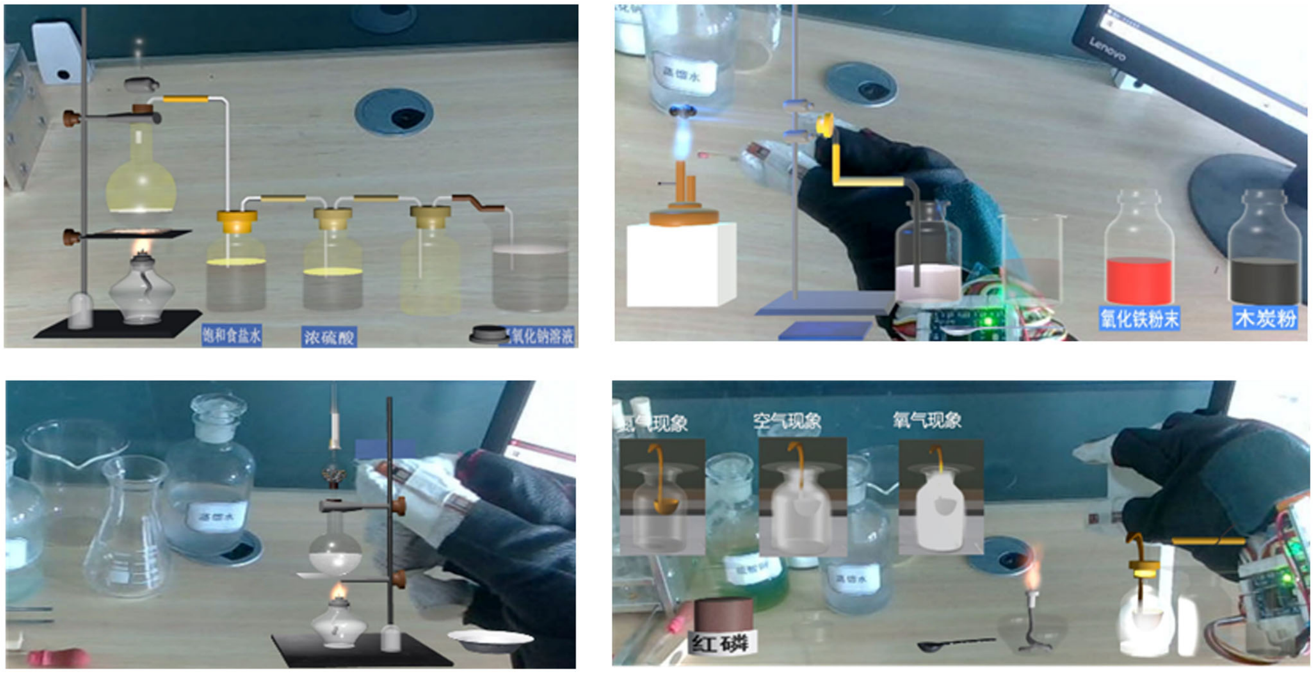

Figure 5. (a) shows the chlorine preparation experiment; (b) the charcoal reduction of iron oxide experiment, the user ignites the alcoholic blowtorch to observe the experimental phenomenon; (c) the ammonia production experiment, the user picks up the red litmus paper turns blue; (d) the red phosphorus combustion experiment, the burning red phosphorus into the gas collection bottle containing oxygen reaction.

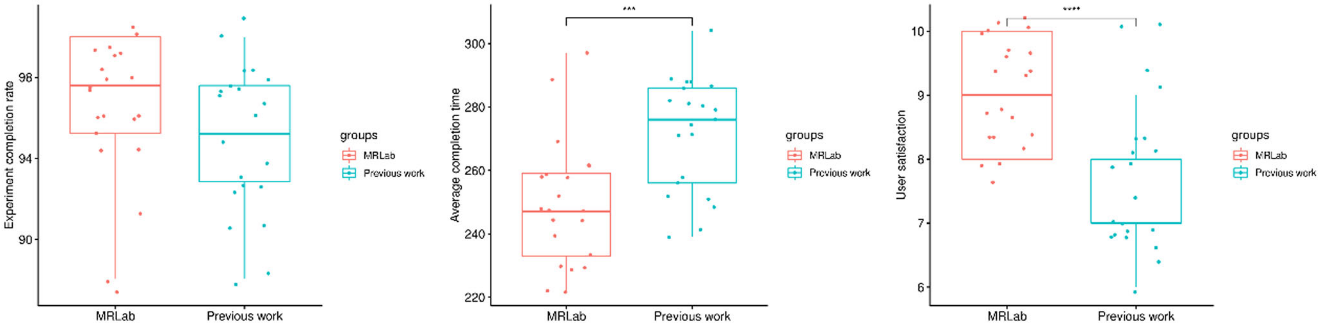

Figure 6. Comparative box plots of experiment completion rate, average completion time and user satisfaction.

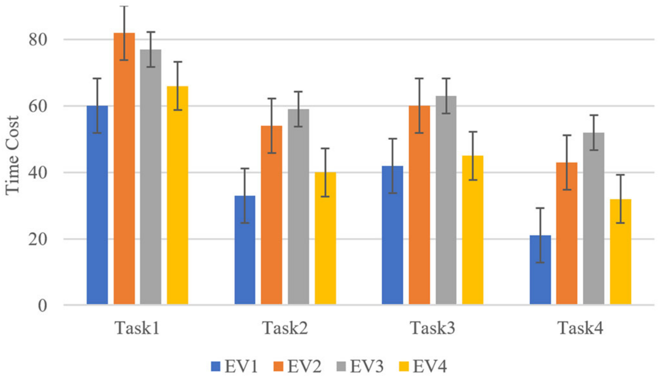

Figure 7. Time performance of four experimental environments on four key tasks (the X-axis represents different key tasks and the Y-axis represents the completion time of that task).

Table 4. The post-hoc comparison test results of interaction time.

	EV4&EV1		EV4&EV2		EV4&EV3	
Task	*p*	S	*p*	S	*p*	S
1	0.0426	Y↑	<0.001	Y↓	<0.001	Y↓
2	0.00242	Y↑	<0.001	Y↓	<0.001	Y↓
3	0.01589	N	<0.001	Y↓	<0.001	Y↓
4	<0.001	Y↑	<0.001	Y↓	<0.001	Y↓

ANOVA and post-hoc comparison tests are performed in this paper to identify significant variations between the experimental groups. The ANOVA results revealed that different environments have a noticeable impact on interaction times for Task 1 ($F = 44.5221$, $p < 0.001$), Task 2 ($F = 72.9653$, $p < 0.001$), Task 3 ($F = 112.745$, $p < 0.001$), and Task 4 ($F = 267.611$, $p < 0.001$). Table 4 reports the comparative test outcomes between EV4 and other experimental settings. In the table, Y and N are used to denote the presence or absence of statistically significant differences.

Based on the graphs and user feedback, EV4 exhibited superior time performance in completing all four tasks compared to the other three experimental environments. In EV1, users perform experiments using only a mouse and keyboard, and they can complete the four experimental tasks in less time if their familiarity with the experimental background. Consequently, EV4 consumed significantly more

Table 5. Questionnaire for user experience evaluation.

Question	Description (low score → high score)
Q1	Whether you are willing to use this experimental platform (unwilling → willing)
Q2	Whether the function of this experiment platform is simple (difficult → simple)
Q3	Whether the interaction process of this experiment platform is simple (difficult → simple)
Q4	Whether you need to ask for help to complete the experiment (required → not required)
Q5	Does it require a lot of learning before using this experiment platform (required → not required)
Q6	Whether this experiment platform is worth promoting (no → yes)
Q7	Whether you can focus more on learning knowledge (no → yes)
Q8	Whether you are more interested in teaching experiments after using it (no → yes)
Q9	Whether I can learn more by using this platform for experiments (no → yes)
Q10	Confident or frustrated in the process of using it (frustrated → confident)
Q11	Is the process of using it psychologically uncomfortable (uncomfortable → comfortable)
Q12	Is it physically uncomfortable during use (uncomfortable → comfortable)
Q13	How do you feel about self-performance using this system (not good → good)

time than EV1 in Task 1, Task 2, and Task 4. In contrast, EV2 and EV3 do not offer the same level of convenience to users, as it requires them to have proper VR experience and operate virtual or real objects within the scene, which increases the time consumption in turn. Conversely, this study proposes the GVVS hybrid fusion model suitable for the EV4 environment, which significantly reduces the time consumption compared to EV2 and EV3, and even for Task 3, the time performance of EV4 is similar to that of EV1.

4.5. User experience evaluation

After users completed Task 1 through Task 4, this study invited volunteers to evaluate four experimental environments based on SUS questionnaires to verify the superiority of EV4's user experience. The questionnaires contained 13 questions with mean opinion scores (MOS) ranging from 1 to 5, symbolizing increasing levels of satisfaction. Table 5 illustrates the questionnaire topics covered in the evaluation. Questions 1 through 6 investigated users' system usage experience ($F = 20.271$, $p < 0.001$), questions 7–9 focused on users' knowledge learning experience ($F = 9.8178$, $p < 0.001$), and questions 10 through 13 delved into users' psychological burden ($F = 6.8708$, $p < 0.001$).

ANOVA and post-hoc comparison tests were used to analyze the significance of differences between groups, based on the questionnaire scoring data. Mean scores along with error bars for system usage experience, knowledge learning experience, and psychological burden for the four experimental settings are illustrated in Figure 8. The comparison tests between EV4 and each comparative experimental setting are presented in Table 6.

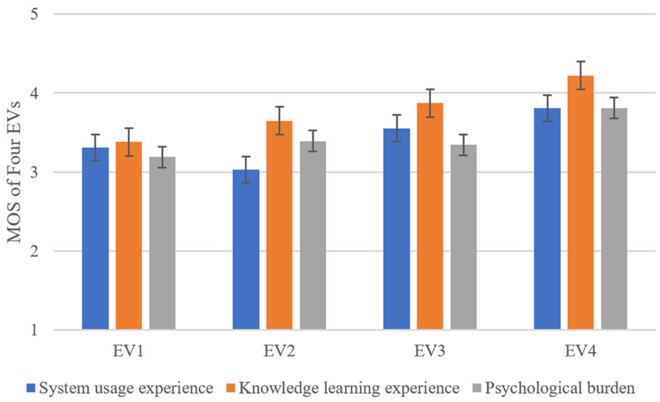

Figure 8. SUS questionnaire survey evaluation of four experimental environments (the X-axis represents the system usage, knowledge learning experience and psychological burden of the four experimental environments, and the Y-axis represents the MOS of different experimental environments).

Table 6. The post-hoc comparison test results of user experience evaluation.

	EV4&EV1		EV4&EV2		EV4&EV3	
Evaluating indicator	*p*	S	*p*	*p*	S	*p*
System usage experience	<0.001	Y↑	<0.001	Y↑	0.009467	Y↑
Knowledge learning experience	<0.001	Y↑	0.000886	Y↑	0.033981	Y↑
Psychological burden	0.000169	Y↑	0.013088	Y↑	0.002801	Y↑

According to user feedback and charts, EV4 offers a notably superior user experience than the other three experimental environments. In EV1, users can only perform experiments via a mouse and keyboard, resulting in limited immersion and operational experience. Furthermore, the absence of voice and text prompts means users have to focus on learning the system, increasing psychological pressure. In contrast, EV2 enables users to perform experiments with VR glasses in a virtual environment, enhancing immersion, while voice prompts help in learning experimental skills. However, using the tactile device can cause physical and mental fatigue, with the entirely virtual environment failing to provide a genuine operational feel. For EV3, although the experiment form is changed from virtual to virtual-reality fusion form, which maintains immersion and improves the sense of operation. Still, additional devices like Kinect and a camera increase the memory burden, making it challenging to operate the experiment. Conversely, EV4 leverages a smart glove and head-mounted device, eliminating the occlusion issue that existed in EV3 and enabling virtual-reality fusion experiments in an MR environment. As a result, EV4 substantially improves system usage experience, knowledge acquisition experience, and psychological burden reduction. Finally, Table 7 depicts a positive correlation between these three indices, indicating high-scoring experimental environments offer a better system usage experience and knowledge acquisition experience with little psychological strain.

4.6. Discussion and future work

The findings of ANOVA and post-hoc comparison tests demonstrates that MRLab (EV4) outperforms the WEB experiment (EV1), AR/VR experiment (EV2), and traditional MR experiment (EV3) in terms of usage experience,

Table 7. Pearson correlation coefficient of different experimental environment evaluation indexes.

Pearson correlation coefficient	EV1	EV2	EV3	EV4
System usage experience and knowledge learning experience	0.62	0.59	0.43	0.43
Knowledge learning experience and Psychological burden	0.76	0.56	0.61	0.76
Psychological burden and knowledge learning experience	0.43	0.69	0.64	0.62

knowledge learning experience, and psychological stress. Furthermore, the correlation analysis indicates that the improvement in one evaluation indicator relates to an improvement in the other two, hence, users will have a better experience and learn more when their psychological stress is minimal. The experimental form of MRLab effectively overcomes the issues associated with (1) EV1's lack of real operating experience and weak immersion, (2) EV2's tendency to cause greater psychological and physical burdens due to prolonged use of HMD devices, and (3) EV3's unsuitability for experimental activities in a pandemic environment because it not only causes users a severe psychological strain but also because its recognition accuracy declines with distance. Consequently, on the basis of EV3, MRLab employs a simple head-mounted device and a smart glove to build a smart experimental environment. It applies the multimodal hybrid fusion model GVVS to analyze and process information such as the user's sensing, voice, and scene vision. After MRLab obtains the user's experimental intention, it can assist the user to realize interactive applications such as the interaction between experimental equipment and chemicals, tactile experience, and observation of experimental phenomena in a virtual-reality fusion environment.

Despite its usefulness as an experimental teaching solution during the pandemic, MRLab has a limited number of common physical and chemistry experiments. Simultaneously, during experiments, many students raised concerns about the difficulty of conducting studies at home due to a lack of instruments. To address this issue, we need to propose new interactive ways to integrate smart experiments into daily life. One possible solution is designing different types of cards to replace various experimental equipment and chemicals, which will assist students in completing experiments. In addition, expanding the types and quantities of experiments can also help overcome these limitations. Regarding wearables, depth information was used in this study to address the positioning issue between the virtual and real coordinates of the smart glove. However, because of depth information errors, the virtual hand's coordinate drift may occasionally occur. To solve this issue, we can employ SLAM technology based on binocular cameras for spatial positioning or upgrade the hardware of smart glove using Hall sensors and AOA of Bluetooth technology (Toasa et al., 2021).

As for potential avenues for future research, we discovered that users occasionally struggle in representing multimodal data in MRLab accurately. To address this issue, we intend to apply the concept of fuzzy categorization to process the user's multimodal input. We can establish a generalized membership paradigm in the MRLab environment to deal with ambiguous multimodal information; Inspired by flow theory (Shernoff et al., 2003), we aim to enhance the user's immersion in the MR experimental environment by introducing personalized game experiments and identifying users' weak knowledge points through application-level innovation. At the same time, we still need to pay attention to human-computer interaction issues, such as interactive, sensory, and learning experiences, and reducing psychological burden, in the virtual-reality fusion environment; In terms of multi-user collaboration, we can establish a new modality for effective remote experimental-teaching interaction between students and teachers. For example, instructors can remotely assist students in MRLab with experimental teaching by using web browsers.

5. Conclusion

The present study employs an MRLab to facilitate a secondary school experiment through the use of a smart glove and a head-mounted device. To infer the user's experimental intention, this study proposes a multimodal hybrid fusion model called GVVS. The GVVS model employs the SrNet model, the SIPA algorithm, and the SVF algorithm to effectively analyze multi-sensor signals, voice data, and scene visual information when the user is performing experimental operations with the smart glove. The model uses hybrid fusion to transform abstract data information into a mathematical language that computers can easily understand, and generates the user's final experimental intention, which is in contrast to multimodal fusion algorithms that only employ feature layer or decision layer fusion. Meanwhile, confronting the issues of a lack of sense of actual operation and immersion in experimental forms like WEB or AR/VR, the real world and the virtual world are subtly displayed in the same visual environment by MRLab, which enhances the user's sense of immersion and realizes the interaction between real and virtual objects. Experiments demonstrate that the GVVS model achieves a good recognition impact of users' experimental intention. During the COVID-19 epidemic, it can assist students in solving the challenge of completing secondary school experiment assignments at home and advance the advancement of smart education.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This paper is supported by the Independent Innovation Team Project of Jinan City (No. 2019GXRC013).

References

Alzubi, T. M., Alzubi, J. A., Singh, A., Alzubi, O. A., & Subramanian, M. (2023). A multimodal human-computer interaction for smart learning system. *International Journal of Human–Computer Interaction*, 1–11. https://doi.org/10.1080/10447318.2023.2206758

Azmandian, M., Hancock, M., Benko, H., Ofek, E., & Wilson, A. D. (2016). *Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences* [Paper presentation]. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 1968–1979). https://doi.org/10.1145/2858036.2858226

Bai, D., Sun, Y., Tao, B., Tong, X., Xu, M., Jiang, G., Chen, B., Cao, Y., Sun, N., & Li, Z. (2022). Improved Single Shot Multibox Detector Target Detection Method based on Deep Feature Fusion. *Concurrency and Computation: Practice and Experience*, *34*(4), e6614. https://doi.org/10.1002/cpe.6614

Beaudouin-Lafon, M. (2000). *Instrumental interaction: An interaction model for designing post-WIMP user interfaces* [Paper presentation]. Proceedings of the 2000 SIGCHI Conference on Human Factors in Computing Systems (pp. 446–453). https://doi.org/10.1145/332040.332473

Bedeian, A. G., & Mossholder, K. W. (2000). On the Use of the Coefficient of Variation as a Measure of Diversity. *Organizational Research Methods*, *3*(3), 285–297. https://doi.org/10.1177/109442810033005

Chen, B., Jiang, R., Kasetkasem, T., & Varshney, P. K. (2004). Channel Aware Decision Fusion in Wireless Sensor Networks. *IEEE Transactions on Signal Processing*, *52*(12), 3454–3458. https://doi.org/10.1109/TSP.2004.837404

Cai, S., Ke, P., Narumi, T., & Zhu, K. (2020). *ThermAirGlove: A pneumatic glove for thermal perception and material identification in virtual reality* [Paper presentation]. Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (pp. 248–257). https://doi.org/10.1109/VR46266.2020.00044

Chen, S.-Y., & Liu, S.-Y. (2020). Using augmented reality to experiment with elements in a chemistry course. *Computers in Human Behavior*, *111*(2), 106418. https://doi.org/10.1016/j.chb.2020.106418

Chen, Z., Wu, R., Guo, S., Liu, X., Fu, H., Jin, X., & Liao, M. (2021). 3D upper body reconstruction with sparse soft sensors. *Soft Robotics*, *8*(2), 226–239. https://doi.org/10.1089/soro.2019.0187

Chi Chung, K., Chen, B. M., Shaoyan, H., Ramakrishnan, V., Chang Dong, C., Yuan, Z., & Jianping, C. (2001). A web-based virtual laboratory on a frequency modulation experiment. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, *31*(3), 295–303. https://doi.org/10.1109/5326.971657

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179–201. https://doi.org/10.3102/00346543068002179

Fang, S., Xue, L., Liang, Y., Wu, J., Ju, C., & Zhao, M. (2020). *NOBOOK VR experiment test* [Paper presentation]. Proceedings of the 2020 International Conference on Virtual Reality and Visualization (pp. 346–347). https://doi.org/10.1109/ICVRV51359.2020.00095

Hammady, R., Ma, M., Strathern, C., & Mohamad, M. (2020). Design and development of a spatial mixed reality touring guide to the Egyptian museum. *Multimedia Tools and Applications*, *79*(5–6), 3465–3494. https://doi.org/10.1007/s11042-019-08026-w

Han, D.-I D., Jung, T., & Tom Dieck, M. C. (2019). Translating tourist requirements into mobile AR application engineering through QFD. *International Journal of Human–Computer Interaction*, *35*(19), 1842–1858. https://doi.org/10.1080/10447318.2019.1574099

Han, R., Feng, Z., Fan, X., Xu, T., Tian, J., & Meng, J. (2020). *A new intelligent VR biological learning system based on natural interaction* [Paper presentation]. Proceedings of the 4th IEEE Information Technology, Networking, Electronic and Automation Control Conference (pp. 175–179). https://doi.org/10.1109/ITNEC48623.2020.9085016

Heun, V., Kasahara, S., & Maes, P. (2013). *Smarter objects: Using AR technology to program physical objects and their interactions* [Paper presentation]. Proceedings of the 2013 CHI Extended Abstracts on Human Factors in Computing Systems (pp. 961–966). https://doi.org/10.1145/2468356.2468528

Hilliges, O., Kim, D., Izadi, S., Weiss, M., & Wilson, A. (2012). *HoloDesk: Direct 3d interactions with a situated see-through display* [Paper presentation]. Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems (pp. 2421–2430). https://doi.org/10.1145/2207676.2208405

Ishii, H., & Ullmer, B. (1997). *Tangible bits: Towards seamless interfaces between people, bits and atoms* [Paper presentation]. Proceedings of the 1997 ACM SIGCHI Conference on Human Factors in Computing Systems (pp. 234–241). https://doi.org/10.1145/258549.258715

Jacob, R. J. K., Girouard, A., Hirshfield, L. M., Horn, M. S., Shaer, O., Solovey, E. T., & Zigelbaum, J. (2008). *Reality-based interaction: A framework for post-WIMP interfaces* [Paper presentation]. Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems (pp. 201–210). https://doi.org/10.1145/1357054.1357089

Kaptelinin, V. J. C. (1996). Activity theory: Implications for human-computer interaction. In B. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction* (pp. 103–116). MIT Press.

Lee, J., Olwal, A., Ishii, H., & Boulanger, C. (2013). *SpaceTop: Integrating 2D and Spatial 3D interactions in a see-through desktop environment* [Paper presentation]. Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems (pp. 189–192). https://doi.org/10.1145/2470654.2470680

Luo, T., Zhang, M., Pan, Z., Li, Z., Cai, N., Miao, J., Chen, Y., & Xu, M. (2020). Dream-experiment: A MR user interface with natural multi-channel interaction for virtual experiments. *IEEE Transactions on Visualization and Computer Graphics*, *26*(12), 3524–3534. https://doi.org/10.1109/TVCG.2020.3023602

Luzhnica, G., Simon, J., Lex, E., & Pammer, V. (2016). *A sliding window approach to natural hand gesture recognition using a custom data glove* [Paper presentation]. Proceedings of the 2016 IEEE Symposium on 3D User Interfaces (pp. 81–90). https://doi.org/10.1109/3DUI.2016.7460035

Ma, Z., Ben-Tzvi, P., & Danoff, J. (2016). Hand rehabilitation learning system with an exoskeleton robotic glove. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *24*(12), 1323–1332. https://doi.org/10.1109/TNSRE.2015.2501748

Mann, S. (1998). Humanistic computing: "WearComp" as a new framework and application for intelligent signal processing. *Proceedings of the IEEE*, *86*(11), 2123–2151. https://doi.org/10.1109/5.726784

McLeod, S. J. S. (2007). Maslow's hierarchy of needs. *Simply Psychology*, 1, 1–18. https://www.simplypsychology.org/maslow.html

Mistry, P., Maes, P., & Chang, L. (2009). *WUW – Wear Ur World: A wearable gestural interface* [Paper presentation]. Proceedings of the 2009 CHI Extended Abstracts on Human Factors in Computing Systems (pp. 4111–4116). https://doi.org/10.1145/1520340.1520626

Nidhyananthan, S. (2018). Human recognition using voice print in LabVIEW. *International Journal of Applied Engineering Research*, *13*(10), 8126–8130.

Ozdemir, D., & Ozturk, F. (2022). The investigation of mobile virtual reality application instructional content in geography education: academic achievement, presence, and student interaction. *International Journal of Human–Computer Interaction*, *38*(16), 1487–1503. https://doi.org/10.1080/10447318.2022.2045070

Pan, Z., Luo, T., Zhang, M., Cai, N., Li, Y., Miao, J., Li, Z., Pan, Z., Shen, Y., & Lu, J. (2022). MagicChem: A MR system based on needs theory for chemical experiments. *Virtual Reality*, *26*(1), 279–294. https://doi.org/10.1007/s10055-021-00560-z

Papakostas, C., Troussas, C., Krouska, A., & Sgouropoulou, C. (2023). Exploring users' behavioral intention to adopt mobile augmented reality in education through an extended technology acceptance model. *International Journal of Human–Computer Interaction*, *39*(6), 1294–1302. https://doi.org/10.1080/10447318.2022.2062551

Park, G., Choi, H., Lee, U., & Chin, S. (2017). Virtual figure model crafting with VR HMD and leap motion. *The Imaging Science Journal*, *65*(6), 358–370. https://doi.org/10.1080/13682199.2017.1355090

Roy, K., Idiwal, D. P., Agrawal, A., & Hazra, B. (2015). *Flex sensor based wearable gloves for robotic gripper control* [Paper presentation]. Proceedings of the 2015 Conference on Advances in Robotics (pp. 1–5). https://doi.org/10.1145/2783449.2783520

Shernoff, D. J., Csikszentmihalyi, M., Shneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, *18*(2), 158–176. https://doi.org/10.1521/scpq.18.2.158.21860

Silva Jennifer, N. A., Southworth, M., Raptis, C., & Silva, J. (2018). Emerging applications of virtual reality in cardiovascular medicine. *JACC. Basic to Translational Science*, *3*(3), 420–430. https://doi.org/10.1016/j.jacbts.2017.11.009

Sun, Q.-S., Zeng, S.-G., Liu, Y., Heng, P.-A., & Xia, D.-S. (2005). A new method of feature fusion and its application in image recognition. *Pattern Recognition*, *38*(12), 2437–2448. https://doi.org/10.1016/j.patcog.2004.12.013

Sun, L., Liu, B., Tao, J., & Lian, Z. (2021). *Multimodal cross- and self-attention network for speech emotion recognition* [Paper presentation]. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 4275–4279). https://doi.org/10.1109/ICASSP39728.2021.9414654

Thevin, L., Briant, C., & Brock, A. M. (2020). X-road: Virtual reality glasses for orientation and mobility training of people with visual impairments. *ACM Transactions on Accessible Computing*, *13*(2), 1–47. https://doi.org/10.1145/3377879

Toasa, F. A., Tello-Oquendo, L., Peńafiel-Ojeda, C. R., & Cuzco, G. (2021). *Experimental demonstration for indoor localization based on AoA of bluetooth 5.1 using software defined radio* [Paper presentation]. Proceedings of the 18th IEEE Annual Consumer Communications & Networking Conference (pp. 1–4). https://doi.org/10.1109/CCNC49032.2021.9369638

Wang, P., Bai, X., Billinghurst, M., Zhang, S., Han, D., Sun, M., Wang, Z., Lv, H., & Han, S. (2020). Haptic feedback helps me? A VR-SAR remote collaborative system with tangible interaction. *International Journal of Human–Computer Interaction*, *36*(13), 1242–1257. https://doi.org/10.1080/10447318.2020.1732140

Wang, H., Feng, Z., Tian, J., & Fan, X. (2022). MFA: A smart glove with multimodal intent sensing capability. *Computational Intelligence and Neuroscience*, *2022*, 3545850. https://doi.org/10.1155/2022/3545850

Wanick, V., Xavier, G., & Ekmekcioglu, E. (2018). *Virtual transcendence experiences: Exploring technical and design challenges in multi-sensory environments* [Paper presentation]. Proceedings of the 10th International Workshop on Immersive Mixed and Virtual Environment Systems (pp. 7–12). https://doi.org/10.1145/3210438.3210444

Wei, J., Wang, X., Peiris, R. L., Choi, Y., Martinez, X. R., Tache, R., Koh, J. T. K. V., Halupka, V., & Cheok, A. D. (2011). *CoDine: An interactive multi-sensory system for remote dining* [Paper presentation]. Proceedings of the 13th International Conference on Ubiquitous Computing (pp. 21–30). https://doi.org/10.1145/2030112.2030116

Wen, F., Zhang, Z., He, T., & Lee, C. (2021). AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nature Communications*, *12*(1), 5378. https://doi.org/10.1038/s41467-021-25637-w

Zeng, B., Feng, Z., Xu, T., Xiao, M., & Han, R. (2020). Research on intelligent experimental equipment and key algorithms based on multimodal fusion perception. *IEEE Access*, *8*, 142507–142520. https://doi.org/10.1109/ACCESS.2020.3013903

Zhang, H., Chen, Z., Guo, S., Lin, J., Shi, Y., Liu, X., & Ma, Y. (2020). *Sensock: 3D foot reconstruction with flexible sensors* [Paper presentation]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu (pp. 1–13). https://doi.org/10.1145/3313831.3376387

Zhang, L., Xie, Y., Xidao, L., & Zhang, X. (2018). *Multi-source heterogeneous data fusion* [Paper presentation]. Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data (pp. 47–51). https://doi.org/10.1109/ICAIBD.2018.8396165

Zhao, N., Yu, L., Geng, Y., & Chen, X. (2007). Support vector machine-based approach to data-layer multi-source ITS data fusion. *Journal of Transportation Systems Engineering and Information Technology*, *7*(2), 32–37. https://doi.org/10.1016/S1570-6672(07)60013-0

Zhou, T., Fu, H., Chen, G., Shen, J., & Shao, L. (2020). Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis. *IEEE Transactions on Medical Imaging*, *39*(9), 2772–2781. https://doi.org/10.1109/TMI.2020.2975344

About the authors

Hongyue Wang is a graduate student at the Department of Computer Science and Technology, University of Jinan. His research interests lie in human-computer interaction, virtual reality and artificial intelligence research in smart education.

Zhiquan Feng is a professor of Computer Science and Technology at University of Jinan. His work explores human-machine interaction and collaboration issues in topics such as smart education, elderly robots, and robotic arms.

Xiaohui Yang is an associate professor of Computer Science and Technology at University of Jinan. His primary research interests lie in the area of Image and Video Processing.

Liran Zhou is a graduate student at the Department of Computer Science and Technology, University of Jinan. Her research interests lie in human-computer interaction and collaboration in elderly care.

Jinglan Tian is a lecturer of Computer Science and Technology at University of Jinan. Her research interests are human computer interaction, human behavior recognition, etc.

Qingbei Guo is an associate professor of Computer Science and Technology at University of Jinan. His research at intersection of pattern recognition and computer vision focuses especially on human computer collaboration.