• Review •

Gesture interaction in virtual reality

Yang LI[1], Jin HUANG[1], Feng TIAN[1,2*], Hong-An WANG[1,2], Guo-Zhong DAI[1]

1. *Beijing Key Laboratory of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing* 100190, *China*

2. *State Key Laboratory of Computer Science, Chinese Academy of Sciences, Beijing* 100190, *China*

*** Corresponding author,** tianfeng@iscas.ac.cn

Received: 19 September 2018 **Accepted:** 21 November 2018

Supported by National Key Research and Development (2016YFB1001405); Frontier Subject Key Research (QYZDY-SSW-JSC041); Chinese Academy of Sciences hundred people, National Natural Science Foundation of China (61572479) project support.

Citation: Yang LI, Jin HUANG, Feng TIAN, Hong-An WANG, Guo-Zhong DAI. Gesture interaction in virtual reality. Virtual Reality & Intelligent Hardware, 2019, 1(1): 84—112
DOI: 10.3724/SP.J.2096-5796.2018.0006

Abstract With the development of virtual reality (VR) and human-computer interaction technology, how to use natural and efficient interaction methods in the virtual environment has become a hot topic of research. Gesture is one of the most important communication methods of human beings, which can effectively express users' demands. In the past few decades, gesture-based interaction has made significant progress. This article focuses on the gesture interaction technology and discusses the definition and classification of gestures, input devices for gesture interaction, and gesture interaction recognition technology. The application of gesture interaction technology in virtual reality is studied, the existing problems in the current gesture interaction are summarized, and the future development is prospected.

Keywords Virtual reality; Gesture interaction; Gesture recognition

1 Introduction

1.1 Virtual reality

Virtual reality (VR), also known as virtual environment, is a computer simulation system that can create and simulate virtual worlds. It provides an immersive experience to the users by generating various types of interactive activities related to visual, auditory, and tactile perceptions. The virtual environment can be a scenery conceived by a designer or a reproduction of the live scene. Professional equipment such as VR headsets and data gloves allow users to sense and control various virtual objects in real time, creating an experience that users cannot obtain in the real world and generating real response or feedback.

VR has three distinct characteristics: interaction, immersion, and imagination. Interaction refers to the natural interaction between the user and the virtual scene. It provides the users with the same feeling as the real world through feedback. Immersion means that the users feel that they are part of the virtual world in the VR scene, as if they are immersed; Imagination refers to the use of multi-dimensional perception information provided by VR scenes to acquire the same feelings as the real world while acquiring the feelings that are not available in the real world.

1.2 Human-computer interaction

Human-computer interaction refers to the research in design, utilization, and implementation of computer systems by human. It is a technology to study about humans, computers, and the interaction between the two. With the development of computing devices and related technologies, the interaction between humans and computers has become a part of daily life including work, shopping, and communication[1]. The graphical user interface based on mouse and keyboard is the commonly used user interface; however, the user interface based on the Window, Icon, Menu, Pointer (WIMP) interface paradigm is limited to two-dimensional (2D) planar objects, which cannot interact with the objects in the three-dimensional (3D) space, reducing the naturalness of the interaction. The interaction between humans communicates through many different aspects including language expression, body movements, and emotional interactions. The new generation of human-computer interaction should have many characteristics such as human-centered, multimodal, and intelligent. Therefore, it is necessary to propose a new interface paradigm, establish a new user interface, and adopt a new interaction method.

In real life, people use their hands to perform operations such as grasp, move, and rotate. In the process of communication, people spontaneously attract the attention of others by hand movements. Gesture is the way people express their will under the influence of consciousness. Therefore, gesture interaction has two functions: one is the movement of the hand, and the other is the specific meaning of the gesture. The study by Karam showed that compared with other parts of the body, the hand as a common communication device is most suitable for application of human-computer interaction[2]. Gesture interaction is intuitive, natural, and flexible. Therefore, it is also very important for some users with physical disabilities, such as visual impairment and hearing impairment, to interact through gestures.

1.3 Gesture interaction technology in virtual reality

Human-computer interaction, as an important supporting technology of VR, provides a variety of interactive modes based on different functions and purposes, enabling people to obtain immersive feelings in 3D virtual scenes. As an emerging technology, VR is a new interactive human-computer interaction interface. Therefore, a new generation interface paradigm Post-WIMP[3] and Non-WIMP[4] has been defined, and the interaction also extends from a mouse- and keyboard-based 2D graphical user interface to a natural user interface or supernatural user interface. Unlike traditional WIMP, in a natural user interface or a supernatural user interface scenario, the user interacts in a 3D environment simply by VR simulation or interactive scenario design, which eliminates waste of resources generated by repetitive manufacturing and controls production costs. With the continuous development of interactive technology, VR has been widely used in many fields such as game entertainment, medical care, and education. In the future, interaction could be like the movie "Player One" in that users could capture people's movements through wearable devices, become avatars, and interact with other people in the virtual world. As an increasing number of companies and scholars invest in the research of gesture interaction technology and related applications, gesture interactions have engaged a large proportion of the field of natural interaction. To simulate daily life operations in a virtual environment, it is necessary to track and recognize the hand movements, to make the computer understand the true intention expressed by gestures, and to perform related tasks. At the same time, in order to make people have a natural interactive experience in the VR process, relevant feedback must be provided to the user.

Gesture is a conscious or unconscious movement of the hand, arm, or limb. The gestures in the interaction process can be distinguished according to different spatiotemporal operation behaviors,

different semantics, different interaction modes, and different interaction ranges. In the early interaction scenarios, the acquisition of gesture signals was mainly based on wearable sensor devices such as data gloves; with the development of mobile devices such as mobile phones and tablets, touch screens became popular; consequently, the acquisition of gesture signals has developed into visual signal acquisition from computer cameras. In recent years, information acquisition based on high-tech equipment such as electromyographic signal acquisition has gradually become a research focus. Owing to the variability and complexity of the gesture, how to use the existing technology to process the input signal collected by the device, how to determine the spatial position and posture of the hand, and how to obtain an accurate recognition result have great impact on the effect of subsequent gesture interactions. Gesture recognition technology is mainly categorized into gesture recognition based on wearable sensor devices, gesture recognition based on touch devices, and gesture recognition based on computer vision. With the development of gesture recognition technology, it has been applied to many aspects of VR. The following sections of this article will discuss the above content in detail.

2 Gesture definition and gesture classification

Gesture is a posture or movement of the user's upper limbs. Through gestures, people can express their interaction intentions and send out corresponding interactive information[5]. Generally, gestures convey information or intentions through physical movements of the face, limbs, or body[6]. In the VR environment, the posture or movement of the users' body is used as input, and the posture information of the user through the interaction device in the physical world is used as system input. In the interaction process, users can not only generate gestures using hand, finger, and palm, but also generate gestures as an extension of the body through tools such as a mouse, pen, and glove. Users can send simple commands to the system through gestures, such as selecting, moving, and deleting, and can also express more complex intents, such as switching the current interactive scene, controlling virtual objects, and performing virtual actions. In different situations, the same gestures may have different meanings due to factors such as culture or geography. The understanding of gesture semantics is closely related to the cultural background, and the gestures can be classified and summarized according to different situations.

2.1 Classification based on spatiotemporal status

According to the spatiotemporal operating state, gestures are generally classified into static gestures and dynamic gestures. A static gesture is a static spatial posture of a finger, palm, or arm at a certain moment. It usually only represents one interactive command and does not contain time series information. A dynamic gesture refers to the posture changes in a finger, palm, or arm over a period of time. Accordingly, it also contains time information[7,8]. The difference between the two is that static gestures only include spatial gestures without causing changes in spatial position, whereas dynamic gestures are characterized by movements in spatial locations over time, such as waving gestures, which are typical dynamic gestures. Dynamic gestures also include conscious dynamic gestures and unconscious dynamic gestures, in which unconscious movements during physical activity are called unconscious dynamic gestures, and gestures for communication purposes are called conscious dynamic gestures[9].

2.2 Classification based on gesture semantics

According to behavioral and cognitive psychology, Pavlocvic et al. categorized gestures into unintentional movements and intentional movements[10]. Unintentional movements mean that the movement of the hand/

arm does not convey any meaningful information, whereas intentional movement is referred to as the gesture. Gestures are divided into manipulative gestures and communicative gestures according to whether they contain communication characters. Manipulative gestures is the action of control the state of the objects through the hand/arm movement, such as moving and rotating objects through the movement of the arm. The communicative gestures can be customized gestures with a specific information function. In a natural interaction situation, they are usually accompanied by verbal communication, mainly including symbol gestures and action gestures. Symbol gestures have a fixed-constraint meaning and can be further divided into referential gestures and modalizing gestures. Action gestures include mimetic gestures that mimics the actual interaction behavior and an deictic gestures with a direction-indicating function.

Similarly, Ottenheimer believes that gestures can be divided into conscious gestures and unconscious gestures, and the conscious gestures are further categorized[11]. Conscious gestures are sub-divided into emblem gestures, affect display gestures, regulator gestures, and illustrator gestures. Emblem gestures are also referred to as quoted gestures. It is a direct conversion of short language communication and the meanings are related to cultural background, such as a waving hand means goodbye. Affect display gestures are gestures that express emotions or a user's intent and are less relevant to cultural contexts. Regulator gestures control the progress of an interaction. Illustrator gestures are a description of the user's language communication state, emphasizing the key parts of the expression, and are related to the communicator's cognition and language skills. McNeill further classified the illustrator gestures into five categories: beat gestures, iconic gestures, deictic gestures, metaphoric gestures, and cohesive gestures. Beat gestures are usually a repetitive action, which have the characteristics of short time and fast speed; iconic gestures describe the image or action through the movement of the hand; deictic gestures indicate the real position or abstract position of the target; metaphoric gestures describe abstract things through simple but meaningful expressions; cohesive gestures transfer interactive tasks[12]. Unconscious gestures, also known as adaptor gestures, mainly refer to unconscious communication habits generated during communication[13].

2.3 Classification based on interaction mode

According to different interaction methods, gestures can be classified into media gestures, direct contact gestures, and non-direct contact gestures depending on the interaction mode. Media gestures are a single-point input device through physical contact, such as a mouse, a joystick, or a trackball, to transmit the obtained data series into the computing system. Direct contact gestures are gesture operations directly on the input device through part of the body or a physical object tool. When the user touches a device such as a screen, the gesture interaction is turned on, and the touch movement is an intermediate state of interaction. Direct contact gestures also include a single touch gesture and a multi-touch gesture. The single touch gesture is a gesture operation provided by the user through a single point of interest using a pointing input device (such as a mouse or pen) or a body part (such as a hand). A pen gesture is a typical single-touch gesture that records the movement trajectory of a pen and uses a symbol or logo for intention expression. Compared to a single touch gesture, a multi-touch gesture is a gesture operation through input devices and multiple regions of interest of the body. Non-direct contact gestures refer to an operation of conducting an interactive command through a body part or a physical object without physical contact with the system. Unlike the direct contact gestures, the non-direct contact gestures do not directly touch any input device or surface during operation, for example vision camera-based gesture interaction[14].

2.4 Classification based on the scope of interaction

According to the different scope of interaction, gestures can be classified into "stroke"gestures and "mid-air" gestures. The former is used mainly on a support surface based on a handwriting gesture of the upper part of the wrist. That is, the user directly uses the hand or tool to move on the support surface and communicate with the computer for information exchange, such as touch screen interactions. The mid-air gesture is based on the user's limbs or interactive devices moving in a space without a supported surface. That is, the user's limbs or interactive tools are not in contact with the computer while interacting in a 3D space. The stroke gesture has a relatively simple interaction mode and low flexibility whereas the mid-air gesture has higher degree of freedom of interaction and more representation modes. Compared with the former, the interaction range of the latter is usually larger, but it lacks capture ability for action details, and as a result, the recognition accuracy of detailed action is insufficient. According to different display modes, mid-air gesture interaction scenes can be divided into 3D virtual scenes based on a traditional desktop display and VR scenes based on a helmet-like device. Although the traditional desktop display can present a 3D virtual scene, the user usually needs to conduct the gesture interaction within a certain distance of the fixed display. Display environments based on helmets or other off-the-table environments allow the user to stay free from the limitations of the device and the environment, which further increases the range of gesture interaction. Of course, when using the helmet display, although the spatial range is increased for user's activity, the user cannot interact on a fixed surface. In this situation, it is an effective choice to use non-contact 3D gestures for interaction.

3 Gesture interaction device in virtual reality

There are a variety of gesture interaction devices used for VR interaction, which enable users to interact more realistically with objects in the virtual world. For different devices, the gesture information input can be classified into wearable-sensor device-based input, touch device-based input, and computer vision interaction device-based input. Wearable sensor-based user interaction devices mainly include data gloves and accelerometers. Touch-based interactive devices mainly include capacitive touch-screen and resistive touch-screen. It is worth mentioning that some devices which have direct contact with the user could have a certain degree of impact on the user's health. For example, mechanical sensor materials may cause allergic symptoms in some users[15]. Computer vision device-based interaction mainly refers to input through the camera, including single camera, binocular camera, and depth camera. This method does not require direct contact, and is more user-friendly, but the configuration and processing method of the device is relatively complicated. There are also many gesture interaction output display devices in VR, such as computer monitors, VR glasses and VR helmets. Users can select appropriate input and output devices according to different interaction scenarios in order to build a gesture interaction system in a VR environment. The following introduces different types of gesture interactive input devices.[1]

3.1 Wearable sensor-based device

3.1.1 Data gloves

As an important collection device for VR gesture interaction, data gloves can collect the posture and motion of the human hand in real time. The device directly detects the activity signal of each joint of the hand through various sensor arrays. It uses the coupling relationship between the data vector of each part

[1] The trademarks of the following products are owned by their respective companies

of the hand, measured by the corresponding sensor, and the data vector of each bending angle of the hand, to obtain the degree of bending or stretching of each part, and to locate the 3D position of the hand. Data gloves can be classified into VR data gloves and force feedback data gloves. The VR data gloves can collect the posture and motion information of the hand, and force feedback data gloves add a force feedback function based on the VR gloves[16]. In the interactive system of VR, the data gloves can achieve two functions. On the one hand, the data glove acts as an input device, and can collect the gesture movements of the user in real time and convert the collected signal into a virtual hand motion. The user can observe the activity of the real hand in the virtual space through the movement of the virtual hand and can operate a virtual target by using various gestures. On the other hand, when using the feedback data glove in the virtual space, the output control of the feedback device enables the user to feel the physical property of the target during the operation and increases the realism for the user. The advantage of the data gloves is that they are not affected by the environment, the amount of data and calculation is small, the recognition accuracy is high, and the 3D information and the movement information of the fingers can be directly collected. The disadvantage is that the manufacturing process of such equipment is relatively complicated, the cost is high, the flexibility is low, and calibration is required frequently.

There are a variety of data gloves available on the market. This article will briefly introduce several classic data gloves[17]. In 1987, VPL produced the first commercial "Data Glove", which uses fiber optic sensors to detect the degree of bending of the fingers. NASA Research Center developed a VIEW system using VPL data gloves and other technologies[18]. 5DT developed a variety of data gloves based on a different number of sensors, mainly divided into five contacts and 14 contacts, both of which use a photoelectric bending sensor and position tracker to detect the spatial motion trajectory of the gesture. The five contacts type has one measuring point for each finger, but two for the 14 contacts type. The advantage of both gloves with five and 14 contacts is that the anti-interference ability is strong, the data quality is high, and the disadvantage is that the force feedback cannot be provided; The "Cyber Touch" developed by Virtual Technology has motion perception and vibration tactile feedback. By installing a small vibrator in the finger and palm area, the precision can reach up to 1°. The "Dexmo" developed by Dexta Robotics uses an exoskeleton link to mimic the bone structure. It collects the hand data through the sensor between the finger joints, and calculates the change of the movement between the joints to estimate the position and posture of the finger. The core benefit is to provide different feedback power to each finger according to the shape and softness of the virtual item, so that the user can experience real feelings in a virtual scene. Figure 1 shows two different types of data gloves, Cyber Touch and Dexmo.

Figure 1 Data gloves. (a) Cyber Touch II[2]. (b) Dexmo[3].

3.1.2 Inertial sensor

The micro-electro-mechanical system (MEMS) is an important branch of miniaturized inertial sensing. It is an independent intelligent system, mainly composed of three parts: sensor, actuator, and micro energy, which can effectively collect information about gestures. Common MEMS systems include MEMS

[2] http://www.cyberglovesystems.com/, product name: Cyber Touch II, company: Virtual Technology

[3] http://dextarobotics.net/, product name: Dexmo, company: Dexta Robotics

accelerometers, MEMS gyroscope sensors, and MEMS pressure sensors. The sensors have high sensitivity and are usually fixed to the wrist, arm, or other positions to obtain gesture data, bringing new tools for gesture interaction. These devices are not affected by the external environment when collecting data and can be nicely applied to scene control in VR. MEMS accelerometers provide acceleration information of object motion, which is widely used in business and scientific research. Rekimoto J. invented the wrist-worn device GestureWrist, which looks like a watch and captures the movement of the hand through an acceleration sensor on the wrist[19]. Baek et al. analyzed the motion of the mobile phone through the amount of accelerometer change, and was able to determine the motion of the user from the posture of the mobile phone[20]. In 2006, Nintendo released the joystick Wii Remote, which looks like a TV remote control. It uses Bluetooth to connect to a computer, adds a 3D accelerometer and an infrared sensor, and can capture the player's arm, wrist, and other gestures to be used in virtual objects in the game. Figure 2 shows the wrist device GestureWrist with acceleration sensor invented by Rekimoto J., and the Nintendo Wii. The MEMS gyro sensors, also known as angular velocimeters, can measure the angular velocity of rotating objects.

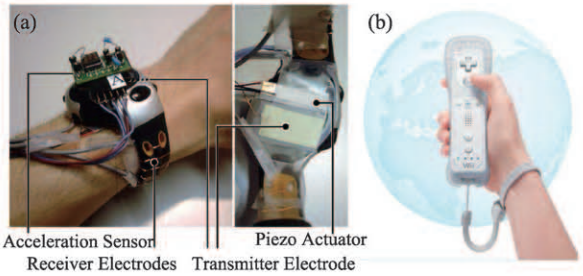

Figure 2 (a) GestureWrist[19]. (b) Wii[4].

3.1.3 Myoelectricity (EMG) sensor

In recent years, myoelectricity is considered as one of the bioelectrical signals. Gesture recognition by collecting myoelectrical information has been widely studied, mainly through recognizing the biopotential changes of the myoelectric signal generated during the movement of the muscle tissue of the hand, to identify the current gesture of the user. This method is not easily affected by the external environment and is especially suitable for people with physical disabilities. In order to obtain accurate bioelectrical information, it is necessary to place at least two or three electrodes on the skin using conductive gels, which may stimulate allergic reactions in people with sensitive skin, and the skin condition, sex, age, and other factors changes affect on the signal generation, therefore, it is difficult to establish a unified measurement standard[21]. Currently, the identification device for myoelectricity detection of gestures is a wristband. In 2013, Thalmic Labs introduced the Myo smart wristband, which captures the bioelectrical change information of the user's arm muscle during exercise through a sensor in the wristband. It can identify 20 different interactive gestures after analyses to judge the intention of the wearer and can create visual control in the virtual environment to implement human-computer interaction. This method has better recognition even with slight movements and is a research focus in the future. Similarly, the wristband Dting developed by BicQ Technology and Innovation Co., Ltd. In Xi'an, China also recognizes the movement position and mode of the finger through the difference between the myoelectrical signals generated by the user when performing different actions. This device provides a friendly human-computer interaction interface in the Chinese language. At the same time, Dting can monitor different aspects of the users' arm including muscle strength, fatigue, and muscle damage. Therefore, it can be used for sign language translation on deaf people, muscle health monitoring in sports medicine, rehabilitation training, and so on. The Myo armband is shown in Figure 3.

[4] http://wii.com, product name: Wii, company: Nintendo

3.2 Touch device

3.2.1 Touch screen

Figure 3 Myo[5].

The touch screen is composed of a touch detection component and a controller. The device detects the touch position of the user through the detection sensor, and then the controller converts the received information into contact coordinates and transmits them to the CPU. The touch screen provides an efficient connection between the human and the computer. Touch screens include capacitive touch screen, resistive touch screen, infrared technology touch screen, surface acoustic wave touch screen, and pressure touch screen[22].

The resistive touch screen causes the resistance of the conductive layer to change after touching at the contact position, and the position of the touched point is calculated by detecting the amount of signal change. The advantage is that it is not affected by dust, moisture, or dirt, and it can adapt to harsh environments with a wide range of application. However, the regular resistive touch screen only supports single touch, and when there is a high external pressure or scratches the touch screen will become damaged, affecting the service life. The capacitive touch screen uses touch to change the contact capacitance, and the touch position is obtained by measuring the frequency change of an oscillator. It has the advantages of long life, easy assembly, and multi-touch support. Most of the multi-touch screens in the market are based on capacitive touch screens, such as the iPhone. Rekimoto J added a layer of electrodes to simulate a capacitive sensor in regular clothes. Wearing such clothing, the user can create an interactive gesture by selecting a specific area thereby sending a specified command to the system[19]. The infrared touch screen uses an infrared transmitting and receiving device to detect the position where the user's contact blocks the infrared signal. The infrared touch screen is not affected by current, voltage or static electricity, and the cost is low, but it is sensitive to the lighting environment. The surface acoustic wave touch screen transmits a high-frequency mechanical wave that propagates along the surface of the medium and determines the position coordinates by blocking the position of the sound wave on the contact. The surface acoustic wave touch screen has a long life and good resolution. It is not affected by environmental factors such as temperature and humidity and is suitable for use in public places. However, the sound wave emission will be affected by factors such as dust and grease; therefore, it is necessary to pay attention to environmental sanitation and to perform frequent maintenance. The pressure touch screen is a multi-touch device, based on a regular touch screen, that senses the increased pressure of the pressure-sensing layer at the point of touch. The finger position is calculated by obtaining the pressing force at different positions on the screen. In addition to the use of pressure sensors, pressure touch can also be achieved in a variety of other ways, such as Apple's 3D Touch through pressure-capacitor technology and Microsoft's multi-touch interactive system PixelSense through infrared sensing technology.

3.2.2 Stylus pen

Stylus pen interaction is another common technique for operating a touch surface. The pen-type touch device mainly includes a touchpad and a stylus pen, and uses the stylus pen to interact on the touch surface

[5] https://www.myo.com, product name: Myo, company: Thalmic Labs

of simulated paper to convert the collected data into horizontal coordinates. Currently, the pressure of the pen tip and the 3D posture information of the pen can also be collected by adding a sensor to the pen. The principle of stylus pen technology is mainly achieved by electric tracking technology, magnetic tracking technology, ultrasonic tracking technology, and ray tracing technology[23]. The electric tracking technology generates a weak current through the emitter electrode in the touchpad. The current returns to the receiving electrode in the touchpad through the hand and the pen and estimates the position of the pen by the center position of the signal intensity. The magnetic tracking technology uses a voltage-applied coil to form a magnetic field in the interaction region and determines the orientation of the pen by coupling the transmitted and received magnetic signals. Ultrasonic tracking technology is an acoustic positioning method. The pen is an ultrasonic transmitter and the ultrasonic receiver is installed at the two ends of the interaction area. The time interval between the emission to the receiving of the sound wave is calculated to determine the distance between the sound source and the target. Infrared signals are commonly used nowadays for sound wave transmission. Ray tracing uses a laser emitter in the pen to project to the writing device. After the light is reflected by the device to the pen, the current position coordinates of the pen are analyzed.

The following are a few common smart pen devices. In 1983, WACOM first used the electromagnetic induction method to manufacture an input pen. Primarily, it is for user CAD design. After years of development, WACOM's input pen can also support pressure sensing technology. The Anoto smart pen jointly developed by Hitachi and Anoto Sweden looks the same as a regular pen but needs to be written on a dedicated Anoto paper. The paper is printed with fine gray dots according to special coding rules. A self-carried micro camera is used to capture the arrangement of the points to analyze the position of the pen tip and then obtain a writing track. The current capacitive pens on the market, such as Apple Pencil, have higher precision and lower latency. Phree, produced by OTM Technologies, installs a laser-scattering instrument on the tip of the pen, which scatters the laser light onto the surface of the object and receives the reflected light. Handwriting is obtained by analyzing the reflected light. The object can be recorded anywhere without requiring any other equipment. These digital pens are close to regular pens in appearance. Figure 4 shows the Apple Pencil produced by Apple and Phree, which is known as a futuristic technology in the digital pen field.

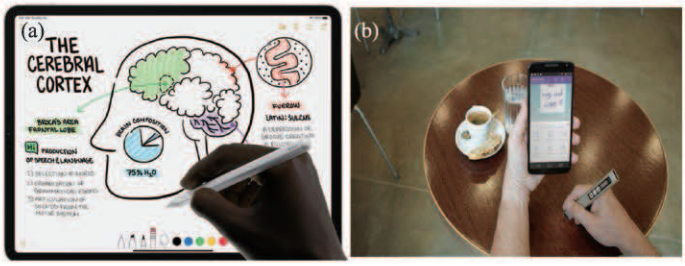

Figure 4 (a) Apple Pencil[6]. (b) Phree[7].

3.3 Computer-based vision interaction equipment

Gesture interaction based on a computer-vision interaction device does not require wearing other external products, and users only need to use their hands within the camera collection range. Its advantages are a low requirement on equipment, low cost, and high collection freedom. The disadvantage is that it is greatly affected by environmental factors, such as light, skin color, and occlusion. The cameras can be black-and-white cameras, color cameras, infrared cameras, and other types. The types of computer vision acquisition devices commonly used in gesture interaction scenarios are mainly classified into monocular cameras, multi-view cameras, and depth cameras.

[6] https://www.apple.com/apple-pencil/, product name: Apple Pencil, company: Apple

[7] http://www.otmtech.com/, product name: Pthree, company: OTM Technologies

The early vision-based gesture interactions used many types of monocular cameras, such as color cameras and infrared cameras. This method obtains relatively simple information, since it is difficult to obtain gesture information in a 3D space. Freeman and Roth used the direction histogram of the image as the feature vector, and through the template-matching method, 10 interactive gestures could be recognized in real time[24]. Zhang et al. used the monocular camera to collect the feature pixel at the edge of the hand area as the recognition feature, and based on the Hausdorff distance-template-matching method, 30 Chinese hand and finger sign letter static gestures were recognized[25]. Meng et al. used regular infrared cameras, collected one to nine numbers and calculated their corresponding grayscale images. Neighborhood transformation and Hausdorff distance were used to segment the gestures, and the gestures were recognized by the support vector machines. Under conditions of arbitrary illumination and complex background, the accuracy of gesture recognition can reach 80%[26]. In addition, PointGrab, EyeSight, and Extreme Reality have developed a gesture control Software Development Kit (SDK), which allows users to perform gesture recognition based on existing regular cameras.

A multi-view camera is a method for simulating binocular- or compound-eye-imaging by two or more cameras. The cameras at different locations are used to collect images of objects at different positions at the same time. The cameras are calibrated to establish a coordinate system to determine internal and external parameters of the camera. The corresponding positions of the pixels formed by the same target are obtained, the corresponding relationship between the features of the images is matched, and the spatial position information of the object is restored. The most common camera for this situation is the binocular camera, which simulates the "parallax" of the human eye and produces a stereoscopic image. The setup only needs to pre-calibrate the position of the camera and use the imaging information of the object in the two images to perform a matching calculation, and to obtain the distance between the object and the camera. For example, Leap Motion, a classic gesture recognition system released by Leap, can reach millimeter-level detection accuracy for the user's hands. Since Leap Motion has higher recognition accuracy and faster processing speed, many scholars use this device for gesture interaction[27], Chen et al. used Leap Motion to capture the motion trajectory of 36 gestures, and used a support vector machine (SVM) to quickly identify and classify gestures[28]. The product Fingo released by uSens Inc. collects images through two built-in infrared cameras. After processing, it can recognize 22 joints of the hands and achieve 26 degrees-of-freedom gesture tracking[29,30]. There are many similar products on the market, such as the zed 2K Stereo Camera from Stereolabs and BumbleBee from Point Grey.

The depth camera calculates the depth information of the object through the depth sensor to avoid the influence of complex or cluttered background on gesture recognition. The depth measurement method is mainly based on light coding and time of flight (ToF). The depth camera made by Microsoft used this depth-measurement method and is widely used in gesture recognition[31,32]. The Kinect 1.0 depth camera acquires the image depth based on the structured optical coding method. Specifically, an infrared emitter is used to optically encode the object in space; the spatial coding result is read by the infrared camera; and the 3D spatial information is obtained through the chip calculation[33]. Kinect 1.0 effectively recognizes distances from 0.8m to 4.0m and uses structured light to capture deep-image products, including Xtion PRO released by ASUS, and the RealSense series released by Intel. The Kinect 2.0 version introduced by Microsoft adopted the ToF method, which improved the depth-recognition effect and the accuracy. The effective recognition distance is increased from 0.5 meters to 4.5 meters. Ren et al. used the color image and depth image of Kinect for gesture recognition[34]. Similarly, Soft Kinetic's ToF-based camera depthSense 541, which was acquired by Sony, provides a very mature hand-interaction device for VR interaction.

There are two major types of hand data collected by the camera: one is marked data, which is marked on positions such as the wrist and fingers to facilitate subsequent modeling[35]. The Sixth Sense gesture interface developed by Mistry et al. at MIT captures the gesture data by marking the fingers with different colored finger sleeves[36]. Although the marked data collection method improves recognition accuracy to a certain extent, it reduces the naturalness of interaction. At the same time, it reduces the accuracy and speed compared with the acquisition method of directly using data gloves and touch screen devices. Another type, the unmarked data collection method, is the most natural form of interaction, but is greatly affected by the environment. Therefore, finding ways to improve the data collection accuracy and to improve the effect of subsequent algorithm processing will be a major research direction in the future (Table 1).

Table 1 Types of camera with classical measurements

Device name	Company	Release time	Distance measured (mm)	Collection method
Leap Motion	Leap	Feb 2013	25–600	Binocular RGB camera
Fingo	uSens	Aug 2016	50–700	Binocular infrared camera
Kinect1.0	Microsoft	Jun 2009	80–400	Based on structured light
Kinect2.0	Microsoft	Oct 2014	50–450	Based on time of flight (ToF)

Compared with the traditional mouse and keyboard input method, gesture interaction in VR requires collecting information from multiple degrees of freedom such as position and direction. Selecting appropriate interactive devices is very important for interaction effects. Although the wearable devices-based interactive method is not easily affected by the external environment during data collection, and the accuracy is high, it requires other devices to be worn, and the materials of some devices may even have adverse effect on the health of the user, which greatly affects the comfort. Gesture interaction based on the touch screen can directly operate on the screen, but it can only interact within the scope of the device, which limits the degree of interaction between the spatial range and the interaction freedom of the user. Gesture interaction based on a computer camera does not require the user to put on an extra devicer. The interaction is more natural, but whether it is a monocular camera, a multi-view camera or a depth camera, the data collected is greatly affected by factors such as lighting changes and complex backgrounds. To ensure the accuracy of subsequent image processing, sometimes the hands of the user need to be marked. Therefore, when performing gesture interactions, the requirements of specific interaction scenarios and tasks need to be considered, and it is necessary to select appropriate devices according to the costs and the technical support capabilities, thereby improving the user interaction experience and interaction efficiency.

4 Gesture recognition technology

The purpose of gesture interaction is to reach the goal of controlling the interaction by tracking and recognizing gestures. Therefore, effectively recognizing the user's gesture has a great impact on the final operation effect. Owing to the randomness and complexity of gestures, gesture recognition techniques are relatively complex. According to different data device collection methods, gesture recognition can be classified into wearable device-based recognition, touch technology-based recognition, and computer vision-based recognition[37].

4.1 Gesture recognition based on wearable device

Gesture recognition based on a wearable device can directly acquire the spatial posture of the hand by using sensors. The data glove is a typical representative of a wearable device input and it is composed of a

number of sensors. During the gesture interaction, the bending posture of each finger can be directly collected and the spatiotemporal parameters of the orientation between the two fingers are processed by data normalization and smoothing. Then, model training for gesture recognition can select effective featured parameters. The collected position parameters generally have higher accuracy. Therefore, the choice of method to train the gesture model has a greater impact on the effect of gesture recognition. Mehdi and Khan used an artificial neural network (ANN) model for sensory glove-based American Sign Language recognition and then translated it into English[38]. Shi et al. obtained the measured values of different output nodes through the sensor and paired the nodes into a back propagation (BP) neural network to recognize the gesture[39]. Liang used the Data Glove and collected 51 basic poses, including six directions, and eight motion primitive data. Then, the hidden Markov model (HMM) was used for modeling, and a vocabulary system that can recognize 250 Taiwanese sign languages was established. The average recognition rate of consecutive sentences composed of these gestures has reached 80.4%[40]. Liu and Lei used Data Glove Ultra 5 for data collection, and combined a BP neural network with Markov to identify Chinese fingerspelling[41]. Wu et al. combined the neural network and the learning decision tree to establish a recognition model and built a Chinese fingerspelling gesture recognition system using the data gloves[42]. Weissmann and Salomon obtained 18 measurements of different finger joints through data gloves, and used BP and radial-basis functions to establish multiple different neural network models for gesture recognition. Some of the neural network models recognize certain gestures and their accuracy rate has even reached 100%[43]. In addition, when the input data of the data glove is converted into a virtual hand model to interact with the object, it is usually necessary to perform collision detection to determine the contact situation between the object and the hand, and through the contact situation it determines the operation condition of the hand. For example, Xu and Li of South China University of Technology used the bounding box method to perform collision detection, and compared the detection results with the gesture library, which determined whether the angle at this time conformed to the definition of the gesture, and thereby implemented the virtual assembly technique of the virtual hand[44]. The accuracy of this method is relatively poor, and it is impossible to implement relatively complex interactions. In the future, it will be possible to determine the relative position between the hand and the object through virtual-force analysis and motion analysis.

With the development of MEMS technology, sensors have become miniaturized and intelligent, which further promotes the development of sensor-based wearable devices for gesture recognition. The spatial position of the gesture can be directly acquired by using sensors for obtaining the angular velocity, acceleration and other motion information of the target, and no gesture segmentation is needed. The interactive gesture can be identified by modeling and analyzing the information. Mirabella et al. created a gesture recognition system that allowed users to navigate digital photos, home TVs or to provide special services to people with physical disabilities through pre-defined sets of gestures. The system uses the acceleration sensor to read the input data of the gesture, trains through the HMM and recognizes the condition of the user-defined gesture, and the user can add a new gesture to the gesture list according to the application needs[45]. Similarly, Kela used the accelerometer controller and HMM to collect and recognize the user's input gestures and to study the effects of gesture modality on user interaction. The results showed that different people prefer different interaction gestures for the same task. For some tasks, gestures are more natural than other interactive modes such as voice, laser pointer, and tablet, and it can even enhance the interaction with other modalities[46]. Xu et al. defined six types of simple gestures by using gyroscopes and accelerometers to acquire gesture data, extracted six characteristics such as gesture length and energy, and used decision tree classifiers to identify the six gestures[47]. He et al. added an accelerometer on a

mobile phone to obtain the motion condition of the gesture when the user performed an operation, extracted three different features of discrete cosine transform (DCT) fast Fourier transform (FFT) and wavelet packet decomposition (WPD), and used SVM to perform classification training. Compared with the other two features, the accuracy of WPD-based feature extraction methods for 17 complex gesture recognitions is relatively high, reaching 87.36%[48]. Henze studied the gesture recognition effect of interactive applications using Wii-controller. The system allows the user to pre-train the operation gestures. Firstly, the original data is smoothed by a filtering method. After smoothing, the training samples are obtained by *k*-mean data compression. Then, HMM training is used to obtain the recognition system, which can fuse the gesture recognition result of this system with other modalities and apply it to multimedia interaction[49].

With the development of physiological computing, an increased amount of physiological information is used in human-computer interaction. The use of bioelectricity for obtaining user's gesture information has become a research focus in recent years. The principle of myoelectricity (electromyography, EMG) for the recognition of user's gestures is to identify the user's current motion based on changes in the muscle signals during exercise. It is a method to directly analyze the user's physiological signals[50,51]. There are two main methods for obtaining muscle signals. One is to implant an electrode under the skin to directly obtain the physiological signals. Although this method is not easily affected by the external environment, it is invasive and has the risk of physical damage to the user. The other common method is to analyze the current state of motion of the user by detecting the change of the electric current on the surface of the skin with placed electrodes. Although this method can be easily affected by the external environment, it is simple, convenient, and less harmful to the user, which complies with the interaction principle of human–computer interaction[52]. There are many ways to use EMG to recognize gestures[53—55]. Saponas studied the perceptual interaction between human muscle activity signals and computational interfaces, where the users did not need to rely on specified actions or other physical devices to interact. Ten sensors were arranged in the narrow-band region near the upper arm of the user. The user's EMG signal was acquired at intervals of 250ms and 74 features including frequency domain power and phase coherence were extracted. The SVM was used to train the classification model. The results showed that this method can be used to differentiate four gesture sets, including finger movement or click, finger position difference, and pressure level[53]. Amma arranged 192 electrodes according to an 8×24 grid with 10mm intervals. The array of aligned electrodes was placed on the forearm muscles to obtain high-density EMG signals. The difference between the electrodes was recorded by two electrodes. One out of eight electrodes did not contain meaningful data, so there were 168 available data channels. The root mean square (RMS) of the extracted channel was used as the input feature, and the Naïve Bayes classifier was used to identify the predefined 27 gestures. In addition, the classifier further analyzed the influence of the number of electrodes and the position of the electrodes on the recognition results[56]. Since the EMG signal for the thumb movement in the forearm is relatively weak, Huang proposed to use a dual-channel mobile phone training-data method to identify the fine-grained gesture of the thumb, including left click, right click, tap, long press, or a more complex finger movement. One of the channels was to use the electrode to obtain the gesture motion of the user, and the other channel was the gesture motion state acquired by the touch screen mobile phone accelerometer as an input signal. Finally, the thumb gesture was classified and recognized by the K-nearest neighbors (KNN) algorithm. This recognition result is of great significance to the thumb-based interactive system[57].

4.2 Gesture recognition based on touch technology

A touch gesture is an operation performed directly on a touch screen with a finger or with other tools.

Based on the touch gesture input method, the touch gesture recognition is mainly classified into single-touch gesture recognition and multi-touch gesture recognition. The Rubine algorithm proposed by Dean Harris Rubine[58] and the $1 algorithm by Wobbrock[59] are two classic single-touch gesture recognition. The Rubine algorithm constructs a gesture feature set by extracting 13 features such as the sine and cosine values of the starting angle and the total length of the gesture for any input gesture, and constructs a classifier based on a statistical recognition method. The accuracy of the classifier for gesture paths that contained 30 single-point locations can reach up to 97%. The key of this algorithm is whether the identification features can be accurately selected. The core idea of the $1 algorithm is to resample the input gesture trajectory, rotate the line connecting the center point of the sampling trajectory and the x-axis to zero degrees, and scale the rotated gesture to the standard square size and perform gesture translation. Then, use the "golden selection search" (GSS) on the processed gestures and search matching template to obtain the best matching score for each template. The gesture with the highest score is the final recognized gesture.

Since the $1 algorithm can only recognize single-stroke gestures, researchers proposed the $N algorithm[60] and $P algorithm[61]. As an extension method of $1, $N can recognize multi-stroke gestures without knowing the direction or the order of the track. The principle is that N gesture tracks can generate 2N cases according to the order and direction. Connect the gesture tracks from head to tail according to the order of gestures and obtain an estimated template set of N gestures. Then, match the gesture track with the template set according to the method of $1, and the highest score gesture is used as the final matching gesture. Owing to the diversity of the arrangement and combination, it could lead to a large $N memory consumption and the execution cost becomes high. Therefore, the $P algorithm proposes to remove the time series information of multiple gestures, turns it into a point cloud set. This transforms the input gesture-matching problem into a point-to-point matching problem of time complexity O(n!), and uses the Hungarian algorithm to reduce the searching time complexity to $O(n^3)$, which can effectively solve the time complexity problem of the $N generation. Currently, many touch interactions are based on multi-touch gesture interaction, such as the multi-finger gesture touch of Apple touchpad. For multi-touch gesture recognition, multi-touch gestures can be pre-processed. Later, a virtual single point is calculated and then recognized using a single touch gesture.

4.3 Gesture recognition based on computer vision

Gesture recognition based on computer vision is the current mainstream identification method. The gesture image information is collected by one or more cameras, and the collected data is pre-processed including noise removal and information enhancement. Then, the segmentation algorithm is used to obtain the target gestures within the image. The current gesture classification and description can be obtained through video processing and analysis, and finally the target gesture is identified by the gesture recognition algorithm. The gesture-based gesture recognition is mainly composed of three parts: gesture segmentation, gesture analysis, and gesture recognition.

The first step is to perform gesture segmentations on the input image. The gesture segmentation process mainly includes two parts: gesture positioning and gesture segmentation. The gesture location process extracts the gesture region from the complex background in the frame sequence of the image that contains the gesture and implements the separation of the gesture from the background. Gesture segmentation segments the current gesture from the background area by using an algorithm after positioning the gesture. Static gestures only require extraction of gesture features for a single frame image, while dynamic gestures

require gesture analysis of the extracted frame sequence. The commonly used gesture segmentation methods include motion information-based detection segmentation, apparent features-based detection segmentation, and multi-mode fusion-based detection segmentation. The motion information-based detection segmentation mainly includes the optical flow method and the difference method. The optical flow method does not need to obtain the image background in advance to represent the gesture motion in a complex environment, but it requires the background image to be still and have high illumination requirements. For example, Hackenberg[62] uses the optical flow method to track gestures in real time. The difference method usually has a better effect on gesture segmentation under static background. For a moving background, the background needs to be modeled and differentiated. For example, Freemanand Weissman used the Running Average method to model the background, and then perform gesture segmentation[63]. In addition, the segmentation methods based on apparent features mainly include skin color segmentation, texture segmentation, hand shape, and contour segmentation. Among them, the skin color segmentation method is the most commonly used method[64,65]. It uses the clustering of skin color in color space to establish a skin color model, such as modeling skin color through RGB color space[66,67], Weng et al. used Bayes to establish the skin color model for gesture segmentation, and then combined the skin color, motion, and contour shape information for gesture recognition, which greatly improved the segmentation accuracy[68]. These segmentation methods are not affected by hand shape, but the segmentation error rate greatly increases for light-caused skin color changes. The segmentation method based on multi-mode fusion is used mainly to overcome the influence of the complex environment on segmentation conditions and to combine various features such as apparent features and motion information. When a single-color wearable device or background is used to simplify the segmentation of the scene, in order to improve the accuracy of the segmentation, sometimes the hand is marked, which greatly affects the naturalness of the interaction, and thus the range of application is limited.

Secondly, the gestures are modeled and analyzed. At present, the modeling methods of gestures mainly include appearance-based gesture modeling and 3D model-based gesture modeling. The appearance-based gesture modeling can be classified into 2D static models and motion models. The commonly used strategies are color characteristics, silhouette geometry, deformable gabarit[69,70], and moving image parameters. Color characteristics are the most commonly used modeling methods. Ren et al. used a single camera to capture the moving images of hands, modeled with various information such as color, motion, and edge, and established a coordinate system for hand movements to recognize gestures[71]. The silhouette geometry based models are established by geometric features such as circumference, centroid, and bounding box. Priyal and Bora used the rotation normalization method to align the image of the gesture area and used the KravtCouk moment as the contour feature training model to identify the static gesture[72]. A deformable gabarit-based models are built through a collection of object contour interpolation nodes that can describe global motion. The simplest interpolation method is a piecewise linear function. For example, Ju et al. used the snake algorithm to analyze the video browsing and indexing gestures during the lecture[73]. Motion image parameters-based models are different from the others in that they are modeled by motion parameters such as translation, rotation, and orientation of objects within the video sequence. Luo et al. proposed to use a new descriptor in the real scene and a local motion histogram to describe the motion pattern, and then select the distinguishing features by the boosting method[74]. The apparent gesture modeling is established only by a few local features, and the computational complexity is relatively low. The 3D-based gesture model is used to establish the 3D model of the current image and calculate the gesture state according to the relevant parameters. The commonly used gesture models include 3D skeleton model, 3D textured volumetric model, and 3D geometric model. The skeleton model is the most commonly

used 3D model. For example, Shotton et al. used the depth data generated by Kinect for bone modeling[75]. Gesture analysis consists of two parts: feature detection and parameter estimation. Feature detection is used to extract image feature parameters from the segmented gestures. Image features mainly include visual features and semantic features. Visual features include color, texture, and contour[76], semantics features represent an understanding of the image content. Parameter estimation uses different methods depending on the model. For example, Lu and first described the athlete's area using the histograms of oriented gradient (HOG), and then used principal component analysis (PCA) to project the HOG into the linear subspace and obtained PCA-HOG description features. Experiments with hockey and soccer showed that this method is robust to track and recognize results under changes in illumination, attitude, and viewpoint[77].

Finally, the recognition of gestures is mainly based on traditional machine learning methods and neural networks. There are many methods to identify vision-based gestures using traditional machine learning models. For example, template recognition is used to identify static gestures, and HMM related to time domain information is used for dynamic gesture recognition[78]. The selected recognition method has great correlation with the type of gesture. Keskin et al. used a pair of common web cameras to capture the user's gesture data while wearing a pair of colored gloves, and trained the HMM model to perform real-time tracking and recognition on the user's eight predefined gestures[79]. In recent years, with the development of deep learning, this technology has been widely applied to the field of gesture recognition. For example, Chai et al. simultaneously collected color images and depth images, and extracted the skeletal features and gradient histogram features of the gestures. Then, they fused the extracted features to establish FAST-RNN to segment the continuous gesture into isolated gestures. by a simple recurrent neural network (SRNN). The dual-flow recurrent neural network 2S-RNN was established for continuous gesture recognition through simple recurrent neural network (SRNN) and Long Short-Term Memory (LSTM)[80]. Tsironi et al. combined the sensitivity of convolutional neural networks (CNN) to visual features and the validity of LSTM to continuous events, and proposed the convolutional long short-term memory recurrent neural network (CNNLSTM)[81].

4.4 Multimodal interaction technology and gesture recognition

In daily communication, people not only use gestures, but also express their information through various sensing methods such as voice, touch, and eye contact. Therefore, multimodal interaction provides two or more modalities collaboration interaction methods. Using the different sensing modalities of the user to interact with the computer is more natural and efficient and is the future way for interaction. For different modalities, the information between the modalities have the characteristics of complementary and redundancy. For example, when the user describes the object in language, he or she will unconsciously swing the arm to enhance or supplement the description. In the VR scene, choosing the appropriate multimodal interaction technology and fusion method is very important to improve the efficiency of interaction and collaboration. The combination of gestures and voices can intuitively express the user's intention. The user can confirm the meaning of the gesture through voice commands, and gestures can eliminate the influence of noise on the operation. Therefore, gestures are often used in VR interaction. The "Put that there" system proposed by Bolt is one of the earliest multimodal interactive systems. This system combines both gesture and voice, enabling users to generate and edit the shape, color, size, and other attributes of a graphic in the large-screen display interface through voice and pointing gestures[82]. The QuickSet system developed by Cohen et al. established a map navigation system that supports military

training by analyzing voice and gesture in real time[83]. In addition to the voice modality, gestures can also be merged with other modalities, such as gaze. The gesture modality is merged with the gaze to perform discrete or continuous interaction on different scenes. For example, when the user performs a task of closing the web window, the icon can be tapped at any position while the close icon is being stared at[84]. When moving the position of the object, it is not necessary to point directly to the object to be moved. The user can select the object to move by using their sight, and then specify the trajectory of the object movement by the gesture trajectory at any starting position. The object selected by sight will follow the gesture trajectory to its final position[85].

The current multimodal fusion method mainly includes early fusion and late fusion. Early fusion refers to the fusion of the original data at the signal level. This fusion method is suitable for tight coupling between modalities. The late fusion uses a unified data representation after the independent processing of data of each modality, and then conducts semantic fusion and analysis. This fusion method is more suitable for multimodal fusion with different time scale features; therefore, it is widely used[86,87]. Sim proposed the SAYS multimodal interactive system, combining the two modalities of voice and gesture keyboard to improve the efficiency and accuracy of text input. HMM was used to train the gesture and the voice models, and then the probabilistic method was used to fuse the models. The fusion model shows that the complementary information provided by the two modalities can effectively distinguish some confusing words, thus improving the accuracy of text input prediction. For an input system of 20000 words, the text prediction accuracy rate with only the gesture keyboard was 92.2%. However, when combined with the two modalities, under normal noise conditions, the text prediction accuracy rate was 96.4%. In a noisy environment, the text prediction accuracy was 94%[88]. The NUMACK navigation system established by Kopp et al., combines voice and gesture, and it answers questions raised by users about campus buildings and their locations. First, the system maps the acquired specific landmark gestures (such as gesture representations and motion trajectories) to the image description feature, then the system translates the acquired voice information into text and uses the sentence planning using descriptions (SPUD) system to analyze the obtained text information. It then extends the existing SPUD system, converts the image description into a part of the syntax tree, and integrates into the syntax tree based on text information construction. Finally, a hierarchical ontology structure based on multimodal fusion is constructed. In this way, the gesture representation enables the user to maintain higher attention during the communication process, and reduces the ambiguity in the gesture information through the accurate expression of the text, thereby obtaining better expression accuracy and ideographic effect[89]. The representative gesture recognition and interaction techniques developed in recent years are listed in Table 2.

In actual operation, suitable interaction devices can be selected based on different interaction scenarios to establish a gesture recognition system. Different gesture recognition systems have different characteristics. A gesture recognition system based on wearable devices, such as data gloves, can directly obtain information that the user is in contact with the virtual world object, and it has a high accuracy. However, the recognition ability is greatly affected by the performance, quantity, and position of the sensors. MEMS further advances the ability of sensors to acquire data, but since it mainly captures motion information of objects, it is mainly used for dynamic gesture recognition. As an important physiological signal, myoelectricity recognizes the current condition of the user through changes in physiological signals during exercise. It is often used in dynamic gesture interaction scenarios. For muscle surface signal recognition methods, the recognition ability is greatly affected by electrode position, quantity, and environment. The external interference for gesture recognition based on touch technology is relatively small. The single-touch recognition effect is relatively good. For multi-touch, it is often converted into a

Table 2 Representative gesture interaction technology in recent years

Author	Year	Type	Input device	Recognition model	Main content
Weissmann et al.	1999	Static	CyberGlove	BP and radial bias function (RBF) neural networks	Use data gloves to obtain 18 measurements and compare the effects of different neural network models on recognition results
Xu et al.	2014	Static	CyberGlove	Oriented bounding box, angle detection	Four major types of mechanical parts are defined and identified, and corresponding decision algorithms are given according to different conditions
Mirabella et al.	2010	Dynamic	MEM accelerometer	HMM	The user's data is read by an accelerometer, and then the HMM is used to train and recognize the user-defined gestures to establish a gesture recognition system.
Xu et al.	2016	Dynamic	MEM gyroscope and accelerometer	Decision tree	Pre-trained gestures not needed, no template matching, gesture recognition by combining decision-tree classifier with posture angle
He et al.	2008	Dynamic	MEMS accelerometer	SVM	Compare the effects of three different inputs, use SVM modeling on gesture recognition results
Henze et al.	2008	Dynamic	Wii controller	HMM	Randomly assign gestures, filter the data, reduce the amount of input data by clustering, and train with HMM
Saponas et al.	2008	Dynamic	EMG	SVM	Using 10 sensors to acquire EMG signals, it can distinguish four data sets, which is the position and pressure of different fingers, click and move
Amma et al.	2015	Dynamic	EMG	Naive Bayes	Study the effect of different electrode numbers and positions on the recognition results, obtain high-density EMG signals, and record the difference between two electrodes
Huang et al.	2015	Dynamic	EMG+MEMS accelerometer	KNN	Use dual devices to obtain dual-observation input signals; combine EMG signals with MEMS signals to identify user-defined gestures
Rubine	1991	Dynamic	stylus/touch pad	Linear classifier	The gesture feature set is created by extracting the geometric features and dynamic features of the stroke, and the gesture is identified by statistical recognition.
Wobbrock et al.	2007	Dynamic	stylus/touch pad	Golden Section Search	Re-sample, rotate, and scale the input points to match the specified template for higher gesture recognition results
Anthony et al.	2010	Dynamic	stylus/touch pad	Golden Section Search	An extension of $1 that automatically recognizes multi-stroke gestures without a direction and an order of the track
Vatavu et al.	2012	Dynamic	stylus/touch pad	Hungarian algorithm	An extension of $N that removes the time series information of multiple gestures, turns them into a point cloud collection, and transforms the input gesture-matching problem into a point-to-point matching problem, reducing the spatiotemporal complexity
Keskin et al.	2003	Dynamic	stereo	HMM	Users need to wear colored gloves, use a pair of regular webcams to collect binocular visual gesture data, and identify and track pre-defined gestures through HMM.
Chai et al.	2017	Dynamic	Kinect	2S-RNN	Using an RGB-D video stream as the collected gesture data, establishing 2S-RNN through SRNN and LSTM for continuous gesture recognition
Tsironi et al.	2016	Dynamic	RGB camera	CNNLSTM	Combined the sensitivity of CNN to visual features, and the context memory of LSTM, for dynamic gesture recognition
Sim	2012	Dynamic	touchscreen/ icrophone	Statistical based multimodal fusion model	The n-gram is used to build the language model, and the acoustic model and gesture model are established by HMM. Then, the multimodal fusion model is obtained by statistical analysis, and the two modalities of voice and gesture are used for text editing.
Stefan Kopp	2004	Dynamic	RGB camera/ microphone	Syntax analysis, Semantic Analysis	Uses SPUD to syntactically analyze the text obtained by speech translation, then converts the specific gesture into the form of syntactic representation, integrates it into the original syntax tree, constructs a topology for ideology, and applys it to address the automatic response system.

single-touch problem for recognition by preprocessing. The computer vision-based gesture recognition includes monocular, binocular, multi-view, and depth camera recognition methods. They mainly include gesture segmentation, gesture feature extraction and analysis, and establishment of a recognition model. For the monocular camera, the amount of information is relatively small, and the recognition speed is fast, but the results can be greatly interfered by the external environment and the accuracy is relatively low. The

binocular camera can use the relative position coordinates of the camera to restore the spatial position of the gesture and establish a 3D gesture model. It has a large recognition range and a high precision. However, the device needs to be calibrated, and the accuracy is limited by the baseline and the resolution. Furthermore, it contains more information and requires higher computing power. The depth camera does not need to be calibrated. The depth information of the gesture can be obtained by using the structured light or the ToF method, but the measurement range is relatively small, and is susceptible to the influence of sunlight. Therefore, it is mainly applied to indoor scenarios. Since the expressed information of a single gesture modality is relatively limited, using the combination of gestures and other modalities not only provide more interactive information, but also make the interaction more natural. It will be the major interaction method of VR in the future. Future research on how to choose the appropriate fusion method for different interaction scenarios and how to establish an effective multimodal fusion model is warranted.

5 Application of gesture interaction system in virtual reality

With the development of gesture interaction and VR technologies, gesture interaction in VR has been widely used in many fields. In the medical field, gesture interaction is of great significance for improving the efficiency of doctors in treating diseases and helping patients to recover. For surgeons, timely access to patient-related medical-imaging data during surgery can help improve surgical efficiency. For example, Ruppert et al. used the Kinect to track and recognize gestures, enabling a contactless gesture control interface that allowed surgeons to flip through patient data at any time during the medical procedure[90]. Similarly, the computer-vision-based gesture recognition system "Gestix" proposed by Wachs captures physician's gestures in real time through a single camera, transforming the movements of physician's gestures into commands for physicians to access electronic medical records at any time (Figure 5)[91].

For patients, Keskin combined gesture recognition technology with a virtual 3D speaker and developed a multi-modal 3D healthcare communication system that provides the ability to communicate with those who are unable to walk freely and require bed rest for rehabilitation. The user only needs to input gestures to the system, and once the system recognizes the gestures, the 3D virtual speaker will respond accordingly to the demand. The system can recognize nine gesture commands such as hunger, thirst, cold, and hot. Figure 6 shows the "I am thirsty" gesture. This system uses a binocular camera for data collection. Users need to wear colored gloves to mark the hand. After segmenting the gestures in the video stream, the continuous hidden Markov model is used to identify the gesture commands[92] . Phelan et al. used the oculus rift head-mounted display, Kinect, and Myo bracelets to construct a prosthetic rehabilitation training prototype system to help amputation patients reduce the training time of using prosthetic limbs. A virtual kitchen was developed using the Unity game engine, and Kinect was used to track the user's body position. The Myo bracelet was

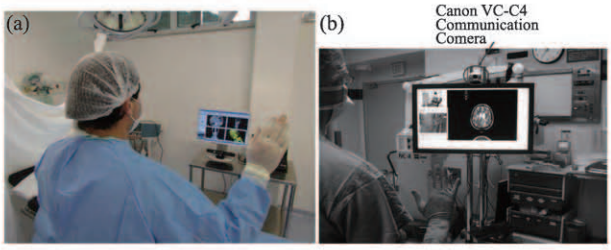

Figure 5 (a) Access to patient data interface using gestures[90]. (b) "Gestix" system configuration[91].

Figure 6 (a) User input gesture "I am thirsty"[92]. (b) 3D virtual speaker[92].

used to detect the muscle activity, and to recognize the user's gesture in order to control the operation of the prosthesis, thereby reducing the service cost of service[93], as shown in Figure 7.

Figure 7 Using prostheses to operate kitchen equipment[93].

In the field of education, gesture interactions are free from the traditional learning and cognitive methods, which increase the interactivity and entertainment in the teaching process. Moustakas et al. used data gloves, Cyber Touch, shutter-type 3D glasses, projection screens and workstations to achieve geometry teaching in VR. The purpose is to add 3D geometric objects to the scene through virtual hands, and solve complex geometric problems, such as geometric problems of 3D European space, by using the relative relationship of object planes, surfaces or spatial positions. The data glove has a force feedback function, and when using a virtual hand to place a geometric object, the collision detection method can be used to calculate the contact level of the object, so that the user can obtain force feedback to increase authenticity. Student test experiences showed that this teaching method is innovative, and the satisfaction with it reached 87%[94]. Figure 8 shows the user solving geometric math problems in the virtual space.

In the junior high school classroom, Kinect was used to track and recognize the gestures, so that the user can control the movement of the geometric figures, and splice the combinations into the targeted shape, thereby completing the specified interactive tasks, as shown in Figure 9. The article compared the gesture-based control method with the traditional mouse control method. The results showed that although the gesture-based target movement control is relatively difficult, this method increased students' interest in learning and participation. Therefore, in the premise of improving the robustness of gesture interaction technology, this teaching method can be promoted as a student's educational tool[95].

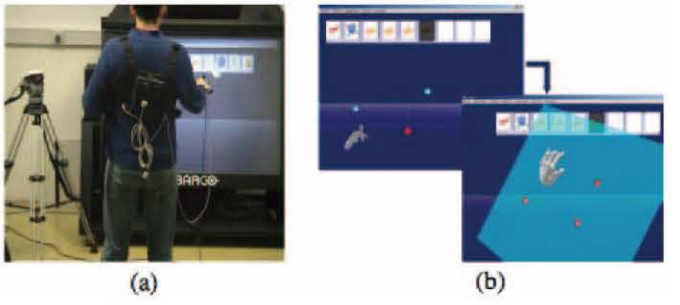

Figure 8 Example of virtual geometry education[94].

Based on the characteristics of information complementary between a three-axis accelerometer and a multi-channel EMG signal, Zhang et al. collected multimodal signals, fused the extracted feature information, and used the decision tree and hidden Markov model to classify and construct a gesture recognition system. This system can recognize 72 Chinese sign language gestures and 40 Chinese hand sentences. It was applied to real-time interactive system control of a virtual Rubik's cube, using 18 kinds of gestures as input commands. It classified and verified the user-related and user-unrelated gestures, and the accuracy of the classification exceeded 90.2%, which confirmed that the interactive framework of

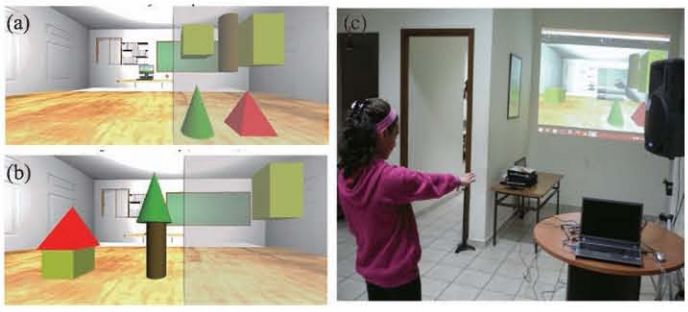

Figure 9 (a) Task initial status[95]. (b) Task completion status[95]. (c) Actual operation scene[95].

this gesture recognition is intelligent and natural[96]. The 18 gestures of the Rubik's Cube operation are shown in Figure 10.

Fan et al. used an Oculus Rift head-mounted display and a Leap Motion device to allow the user to control the user's motion with both hands while roaming in a virtual scene through a virtual avatar (first perspective), using motions such as turning the palm of the left hand up or down to control the user's direction to move forward or backward, while the right thumb is pointing in the direction of the virtual person's rotation, as shown in Figure 11. Experiments showed that the two-hand interaction meets people's daily interaction habits, and increases the immersion of the users into the VR environment. At the same time, motion sickness in the VR environment is reduced. This interactive method is suitable for exploring games such as decryption that do not require quick response speed[97]. Similarly, Khundam used the Oculus Rift head-mounted display to track the user's head position, thereby changing the direction of the viewpoint, and used the Leap Motion device for gesture recognition to control roaming in the VR environment[98]. Park conducted user experience research on four different types of VR games based on gesture interaction, and obtained feedback results of gesture interaction user satisfaction under different visual conditions[99].

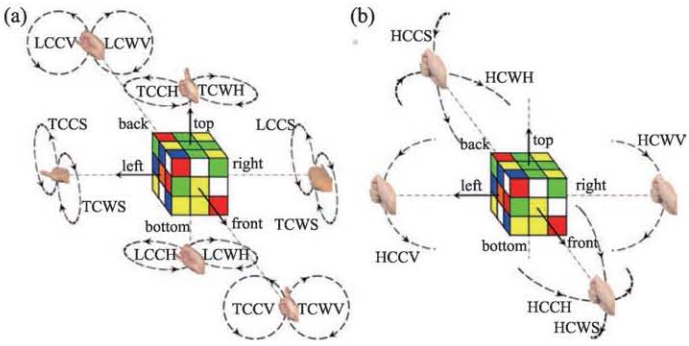

Figure 10 (a) 12 rotation gestures control the rotation of the six sides of the Rubik's Cube[96]. (b) Six rotation gestures to turn the entire cube[96].

Figure 11 Explore the decryption game with two-handed interaction[97].

Gesture interaction alone can only express the user's intention with the hand movement and lacks the interaction information of other modalities. Therefore, combining the gesture movement with other interaction modes to form a multimodal interaction mode is the major research goal in the future VR field. For example, in VR, a voice modality and a gesture modality can be combined to achieve multimodal interaction, as shown in Figure 12. Latoschik used voice and gesture control to connect two objects, such as sending a voice command "Connect two gray sticks with this thing", simultaneously accompanied by the movement of the pointing gesture[100]. Similarly, Chun combined the voice with the gesture, using voice to send operation commands, and performing speech recognition on the

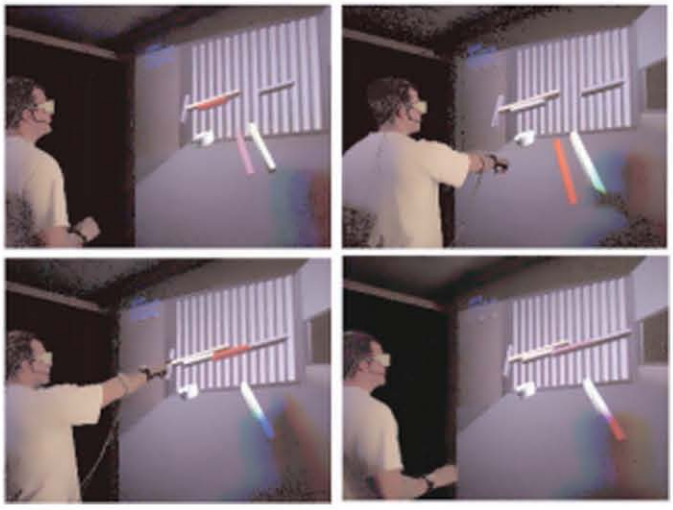

Figure 12 Voice and gesture combined to complete object connection[100].

transmitted voice command. At the same time, Kinect was used to recognize the posture and orientation of the gesture, and collision detection was performed on the virtual hand and the virtual object. Thus, the purpose of moving only virtual objects that meet the virtual gestures was achieved. Multimodal instructions were recognized by voice recognition and gesture recognition, and interactive actions such as selecting, moving, zooming, deleting, and changing outline colors were performed[101].

6 Potential issues

There is still a certain difference between the existing gesture interaction technology and the desired completely natural and efficient interaction. The following aspects can be improved.

Firstly, although gesture interaction simplifies the interactive input method, there is no standardized operation specification. Since the gesture and the task do not have one-to-one correspondence, it is necessary to select appropriate gesture types according to the characteristics of the task in the VR. Owing to the diversity and complexity of gestures, it is difficult for developers to build a consistent operation platform. Therefore, when users use gesture interactions for different products, they need to be trained for a period of time, which increases the difficulty of learning and cognition.

Secondly, gesture-based interaction is closer to the human expression, but in some VR interaction scenarios, it is required to wear cumbersome gesture collection devices, which reduces the comfort and naturalness of the interaction, immersing the users in a negative situation, and affecting their mood.

Thirdly, compared with the point-to-point precision operation of the mouse and keyboard, the gesture interaction is not an accurate operation, and its application range is affected by many factors such as the interaction device, the recognition method, and user proficiency. At the same time, since the gesture interaction recognition process requires a certain processing time, and the VR system itself also has a certain delay, the fluency of the interaction is greatly affected and is not suitable for interactive tasks with high-accuracy requirements.

7 Summary and future directions

This paper summarized the following aspects of gesture interaction technology in VR.

Firstly, the definition of gestures is provided, and the existing gesture classification methods are investigated and summarized. Our study showed that gestures, as a common communication method in human life, can be classified from many different aspects. These classification methods correspond to different interaction scenarios and conditions. Researchers can select an appropriate classification and expression according to the goals that need to be achieved.

Secondly, several VR gesture interaction devices are introduced. They are classified into wearable interaction devices, touch screen-based interaction devices and computer-vision-based interaction devices. Of these, the wearable interaction devices, such as data gloves and acceleration sensors, collect gesture signals, and can directly collect the position information of user's hand in the 3D space. For touch devices, with the support of the multi-touch technology in mobile devices, such as in the iPhone and tablet, the application range is extensive. Compared with other interactive devices, the computer-vision-based gesture interaction method does not require extra tools and is the most natural human–computer interaction gesture acquisition device.

Thirdly, gesture recognition methods are summarized. The gesture recognition result has a great impact on the final gesture interaction. The existing gesture recognitions mainly include wearable gesture

recognition, touch-screen-based gesture recognition and computer-vision-based gesture recognition. The main process of multiple recognition is to pre-process the collected gesture data; to obtain the processed gesture data; to extract the gesture features for modeling and analyses; and finally, to obtain the gesture recognition result. Here, for different input data types, the appropriate modeling method should be selected. The major modeling methods are based on traditional machine-learning models and deep neural networks. The combination of voice, video, and other modal information with gesture information can better identify and understand the current user's intention expressed in the interaction process. This is the focus of current research and the trend of future research.

Finally, the application of gesture interaction in multiple fields of VR is introduced. The combination of gesture interaction and VR technology has greatly changed our lives in many aspects such as medical care, education, daily work, and life. The development of VR provides us with more possibilities to experience feelings that may occur in real life or that cannot even be experienced in real life. In the future, VR technology may subvert our current level of experience and bring us more surprises. At the same time, with the advancement of gesture interaction technology, higher precision, and more natural interaction technologies will be produced. Therefore, applying gesture interaction to VR will bring us a new experience so that we will feel that we are communicating with real "people", and further increase the authenticity, immersion, and interactivity of gesture interaction in VR.

References

1 Pantic M, Nijholt A, Pentland A, Huang T S. Human-Centred Intelligent Human Computer Interaction (HCI²): how far are we from attaining it? International Journal of Autonomous and Adaptive Communications Systems, 2008, 1(2): 168–187
DOI:10.1504/IJAACS.2008.019799

2 Karam M. A framework for research and design of gesture-based human-computer interactions. Doctoral. University of Southampton, 2006

3 Dam A V. Post-WIMP user interfaces. Communications of the ACM, 1997, 40(2): 63–67
DOI:10.1145/253671.253708

4 Green M, Jacob R. Software architectures and metaphors for non-wimp user interfaces. ACM SIGGRAPH Computer Graphics, 1991, 25(3): 229–235
DOI:10.1145/126640.126677

5 Mitra S, Acharya T. Gesture Recognition: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007, 37(3): 311–324
DOI:10.1109/TSMCC.2007.893280

6 Cerney M M, Vance J M. Gesture recognition in virtual environments: a review and framework for future development. Iowa State University Human Computer Interaction Technical Report ISU-HCI-2005-01. 2005

7 Parvini F, Shahabi C. An algorithmic approach for static and dynamic gesture recognition utilising mechanical and biomechanical characteristics. International Journal of Bioinformatics Research and Applications, 2007, 3(1): 4–23
DOI:10.1504/ijbra.2007.011832

8 Rogers Y, Sharp H, Preece J. Interaction design: beyond human-computer interaction. John Wiley & Sons, 2011

9 Rautaray S S, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review, 2015, 43(1): 1–54
DOI:10.1007/s10462-012-9356-9

10 Pavlovic V, Sharma R, Huang T S. Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1997, 19(7): 677–695
DOI:10.1109/34.598226

11 Ottenheimer H J. The anthropology of language: an introduction to linguistic anthropology. Wadsworth Publishing. 2005

12 McNeill D. Hand and mind: What gestures reveal about thought. Chicago, USA: University of Chicago Press. 1992

13 Kanniche M B. Gesture recognition from video sequences. PhD Thesis, University of Nice. 2009

14 International standards: ISO/IEC 30113-11: 2017(E)

15 Nishikawa A, Hosoi T, Koara K, Negoro D, Hikita A, Asano S, Kakutani H, Miyazaki F, Sekimoto M, Yasui M, Miyake Y, Takiguchi S, Monden M. FAce MOUSe: A novel human-machine interface for controlling the position of a laparoscope. IEEE Transactions on Robotics and Automation, 2003, 19(5): 825−841
DOI:10.1109/TRA.2003.817093

16 Tarchanidis K N, Lygouras J N. Data glove with a force sensor. IEEE Transactions on Instrumentation and Measurement, 2003, 52(3): 984−989
DOI:10.1109/TIM.2003.809484

17 Temoche P, Esmitt R J, Rodríguez O. A low-cost data glove for virtual reality. Xi International Congress of Numerical Methods in Engineering and Applied Sciences. 2012, TCG31−36

18 Furness T A. The Super Cockpit and its Human Factors Challenges. Proceedings of the Human Factors Society Annual Meeting, 1986, 30(1): 48−52
DOI:10.1177/154193128603000112

19 Rekimoto J. GestureWrist and GesturePad: unobtrusive wearable interaction devices. In: Proceedings Fifth International Symposium on Wearable Computers, 2001, 21−27
DOI:10.1109/ISWC.2001.962092

20 Baek J, Jang I J, Park K, Kang H S, Yun B J. Human Computer Interaction for the Accelerometer-Based Mobile Game. In: Embedded and Ubiquitous Computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, 509−518
DOI:10.1007/11802167_52

21 Webster J G. Medical Instrumentation-Application and Design. 1978, 3(3): 306

22 Jain A, Bhargava D B, Rajput A. Touch-screen technology. International Journal of Advanced Research in Computer Science and Electronics Engineering, 2013, 2(1)

23 Subrahmonia J, Zimmerman T. Pen computing: challenges and applications. In: Proceedings 15th International Conference on Pattern Recognition ICPR-2000. Barcelona, Spain, 2000, 60−66
DOI:10.1109/ICPR.2000.906018

24 Freeman W T Roth M. Orientation histograms for hand gesture recognition. In: International workshop on automatic face and gesture recognition. 1995, 12: 296−301

25 Zhang L G, Wu J Q, Gao W, Yao H X. Hand gesture recognition based on Hausdorff Distance. Journal of Image and Graphics, 2002, 7(11): 1144−1150
DOI:10.11834/jig.2002011341

26 Meng C N, Lv J P, Chen X H. Gesture recognition based on universal infrared camera. Computer Engineering and Applications. 2015, 51(16): 17−22

27 Weichert F, Bachmann D, Rudak B, Fisseler D. Analysis of the Accuracy and Robustness of the Leap Motion Controller. Sensors, 2013, 13(5): 6380−6393
DOI:10.3390/s130506380

28 Chen Y, Ding Z, Chen Y, Wu X. Rapid recognition of dynamic hand gestures using leap motion. In: 2015 IEEE International Conference on Information and Automation. 2015, 1419−1424
DOI:10.1109/ICInfA.2015.7279509

29 Anonymous. USens CTO Detailed Human-Computer Interactive Tracking Technology. Computer & Telecommunication, 2016(4): 12−13

30 Anonymous. Virtual Reality Technology Trends to "Bare Hand Manipulation". Machine Tool & Hydraulics, 2017(8): 38

31 Gu Y, Do H, Ou Y, Sheng W. Human gesture recognition through a Kinect sensor. In: 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO). 2012, 1379−1384
DOI:10.1109/ROBIO.2012.6491161

32 Ren Z, Yuan J, Meng J, Zhang Z. Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. IEEE

Transactions on Multimedia. 2013, 15(5): 1110–1120
DOI:10.1109/TMM.2013.2246148

33 Raheja J L, Chaudhary A, Singal K. Tracking of Fingertips and Centers of Palm Using KINECT. In: 2011 Third International Conference on Computational Intelligence, Modelling & Simulation, 2011, 248–252
DOI:10.1109/CIMSim.2011.51

34 Ren Z, Meng J, Yuan J, Zhang Z. Robust hand gesture recognition with kinect sensor. In: Proceedings of the 19th ACM international conference on Multimedia. Scottsdale, Arizona, USA, ACM, 2011: 759–760
DOI:10.1145/2072298.2072443

35 Han Y. A low-cost visual motion data glove as an input device to interpret human hand gestures. IEEE Transactions on Consumer Electronics. 2010, 56(2): 501–509
DOI:10.1109/TCE.2010.5505962

36 Mistry P, Maes P, Chang L. WUW–wear Ur world: a wearable gestural interface. In: CHI '09 Extended Abstracts on Human Factors in Computing Systems. Boston. MA, USA, ACM, 2009: 4111–4116
DOI:10.1145/1520340.1520626

37 Rautaray S S, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review, 2015, 43(1): 1–54
DOI:10.1007/s10462-012-9356-9

38 Mehdi S A, Khan Y N. Sign language recognition using sensor gloves. In: Proceedings of the 9th International Conference on Neural Information Processing. 2002, 2204–2206
DOI:10.1109/ICONIP.2002.1201884

39 Shi J F, Chen Y, Zhao H M. Node-Pair BP Network Based Gesture Recognition by Data Glove. System Simulation Technology, 2008, 4(3): 154–157
DOI:10.3969/j.issn.1673-1964.2008.03.003

40 Rung-Huei L, Ming O. A real-time continuous gesture recognition system for sign language. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. 1998, 558–567
DOI:10.1109/AFGR.1998.671007

41 Liu M T, Lei Y. Chinese finger Alphabet flow recognition system based on data glove. Computer Engineering, 2011, 37 (22): 168–170

42 Wu J Q, Gao W, Chen L X. A system recognizing Chinese finger-spelling alphabets based on data-glove input. Pattern Recognition and Artificial Intelligence, 1999(1): 74–78

43 Weissmann J, Salomon R. Gesture recognition for virtual reality applications using data gloves and neural networks. In: IJCNN'99 International Joint Conference on Neural Networks Proceedings. 1999, 2043–2046
DOI:10.1109/IJCNN.1999.832699

44 Xu Y H, Li J R. Research and implementation of virtual hand interaction in virtual mechanical assembly. Machinery Design & Manufacture, 2014(5): 262–266

45 Mirabella O, Brischetto M, Mastroeni G. MEMS based gesture recognition. In: 3rd International Conference on Human System Interaction. 2010, 599–604
DOI:10.1109/HSI.2010.5514506

46 Kela J, Korpipää P, Mäntyjärvi J, Kallio S, Savino G, Jozzo L, Marca D. Accelerometer-based gesture control for a design environment. Personal & Ubiquitous Computing, 2006, 10(5): 285–299
DOI:10.1007/s00779-005-0033-8

47 Xu J, Liu C H, Meng Y X. Gesture recognition base on wearable controller. Application of Electronic Technique, 2016, 42(7): 68–71

48 He Z Y, Jin L W, Zhen L X, Huang J C. Gesture recognition based on 3D accelerometer for cell phones interaction. In: APCCAS2008 - 2008IEEE Asia Pacific Conference on Circuits and Systems. 2008, 217–220
DOI:10.1109/APCCAS.2008.4745999

49 Schlömer T, Poppinga B, Henze N, Boll S. Gesture recognition with a Wii controller. In: Proceedings of the 2nd

international conference on Tangible and embedded interaction. Bonn, Germany, ACM, 2008: 11–14
DOI:10.1145/1347390.1347395

50 Du Y, Jin W, Wei W, Hu Y, Geng W. Surface EMG-Based Inter-Session Gesture Recognition Enhanced by Deep Domain Adaptation. 2017, 17(3): 458
DOI:10.3390/s17030458

51 Kim J, Mastnik S, André E. EMG-based hand gesture recognition for realtime biosignal interfacing. In: Proceedings of the 13th international conference on Intelligent user interfaces. Gran Canaria, Spain, ACM, 2008: 30–39
DOI:10.1145/1378773.1378778

52 Madhavan G. Electromyography: physiology, engineering and non-invasive applications. Annals of Biomedical Engineering, 2005, 33(11): 1671

53 Saponas T S, Tan D S, Morris D, Balakrishnan R. Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Florence, Italy, ACM, 2008: 515–524
DOI:10.1145/1357054.1357138

54 Saponas T S, Tan D S, Morris D, Balakrishnan R, Turner J, Landay J A. Enabling always-available input with muscle-computer interfaces. In: Proceedings of the 22nd annual ACM symposium on User interface software and technology. Victoria, BC, Canada, ACM, 2009: 167–176
DOI:10.1145/1622176.1622208

55 Saponas T S, Tan D S, Morris D, Turner J, Landay J A. Making muscle-computer interfaces more practical. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta, Georgia, USA, ACM, 2010: 851–854
DOI:10.1145/1753326.1753451

56 Amma C, Krings T, Böer J, Schultz T. Advancing Muscle-Computer Interfaces with High-Density Electromyography. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. Seoul, Republic of Korea, ACM, 2015: 929–938
DOI:10.1145/2702123.2702501

57 Huang D, Zhang X, Saponas T S, Fogarty J, Gollakota S. Leveraging Dual-Observable Input for Fine-Grained Thumb Interaction Using Forearm EMG. In: Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology. Charlotte, NC, USA, ACM, 2015: 523–528
DOI:10.1145/2807442.2807506

58 Rubine D H. The automatic recognition of gestures. Carnegie Mellon University, 1992

59 Wobbrock J O, Wilson A D, Li Y. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In: Proceedings of the 20th annual ACM symposium on User interface software and technology. Newport, Rhode Island, USA, ACM, 2007: 159–168
DOI:10.1145/1294211.1294238

60 Anthony L, Wobbrock J O. A lightweight multistroke recognizer for user interface prototypes. In: Proceedings of Graphics Interface 2010. Ottawa, Ontario, Canada, Canadian Information Processing Society, 2010: 245–252

61 VatavuR-D, Anthony L, Wobbrock J O. Gestures as point clouds: a $P recognizer for user interface prototypes. In: Proceedings of the 14th ACM international conference on Multimodal interaction. Santa Monica, California, USA, ACM, 2012: 273–280
DOI:10.1145/2388676.2388732

62 Hackenberg G, McCall R, Broll W. Lightweight palm and finger tracking for real-time 3D gesture control. In: 2011 IEEE Virtual Reality Conference. 2011, 19–26
DOI:10.1109/VR.2011.5759431

63 Freeman W T, Weissman C D. Television control by hand gestures. International Workshop on Automatic Face & Gesture Recognition, 1995: 179–183

64 Kaufmann B, Louchet J, Lutton E. Hand Posture Recognition Using Real-Time Artificial Evolution. In: Applications of

Evolutionary Computation. Berlin, Heidelberg, Springer Berlin Heidelberg, 2010, 251–260
DOI:10.1007/978-3-642-12239-2_26

65 Flasiński M, Myśliński S. On the use of graph parsing for recognition of isolated hand postures of Polish Sign Language. Pattern Recognition, 2010, 43(6): 2249–2264
DOI:10.1016/j.patcog.2010.01.004

66 Bergh M V d, Gool L V. Combining RGB and ToF cameras for real-time 3D hand gesture interaction. In: Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV). IEEE Computer Society, 2011: 66–72
DOI:10.1109/WACV.2011.5711485

67 Jones M J, Rehg J M. Statistical Color Models with Application to Skin Detection. International Journal of Computer Vision, 2002, 46(1): 81–96
DOI:10.1023/A:1013200319198

68 Weng C, Li Y, Zhang M, Guo K, Tang X, Pan Z. Robust Hand Posture Recognition Integrating Multi-cue Hand Tracking. In: Entertainment for Education Digital Techniques and Systems. Berlin, Heidelberg, Springer Berlin Heidelberg, 2010, 497–508
DOI:10.1007/978-3-642-14533-9_51

69 Ju S X, Black M J, Yacoob Y. Cardboard People: A Parameterized Model of Articulated Image Motion. In: Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition. IEEE Computer Society, 1996: 38

70 Kervrann C, Heitz F. Learning structure and deformation modes of nonrigid objects in long image sequences. 1995

71 Ren H B, Xu G H, Lin X Y. Hand gesture recognition based on characteristic curves. Journal of Software, 2002, 13(5): 987–993

72 Priyal P S, Bora P K. A robust static hand gesture recognition system using geometry based normalizations and Krawtchouk moments. Pattern Recognition, 2013, 46(8): 2202–2219
DOI:10.1016/j.patcog.2013.01.033

73 Ju S X, Black M J, Minneman S, Kimber D. Analysis of Gesture and Action in Technical Talks for Video Indexing. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 1997: 595

74 Luo Q, Kong X, Zeng G, Fan J. Human action detection via boosted local motion histograms. Machine Vision and Applications, 2010, 21(3): 377–389
DOI:10.1007/s00138-008-0168-5

75 Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A. Real-time human pose recognition in parts from single depth images. In: CVPR 2011, 2011, 1297–1304
DOI:10.1109/CVPR.2011.5995316

76 Kass M, Witkin A, Terzopoulos D. Snakes. Active contour models. International Journal of Computer Vision, 1988, 1 (4): 321–331
DOI:10.1007/BF00133570

77 Lu W-L, Little J J. Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor. In: Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision. IEEE Computer Society, 2006: 6
DOI:10.1109/CRV.2006.66

78 Moni M A, Ali A B M S. HMM based hand gesture recognition. A review on techniques and approaches. In: 2009 2nd IEEE International Conference on Computer Science and Information Technology. 2009, 433–437
DOI:10.1109/ICCSIT.2009.5234536

79 Keskin C, Erkan A, Akarun L. Real time hand tracking and 3D gesture recognition for interactive interfaces using HMM. In: Proceedings of International Conference on Artificial Neural Networks. 2003

80 Chai X, Liu Z, Yin F, Liu Z, Chen X. Two streams Recurrent Neural Networks for Large-Scale Continuous Gesture Recognition. In: 2016 23rd International Conference on Pattern Recognition, 2016, 31–36
DOI:10.1109/ICPR.2016.7899603

81 Tsironi E, Barros P, Wermter S. Gesture recognition with a convolutional long short-term memory recurrent neural network. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and

Machine Learning (ESANN). Bruges, Belgium, 2016, 213–218

82 Bolt R A. "Put-that-there": Voice and gesture at the graphics interface. Acm Siggraph Computer Graphics, 1980, 14(3): 262–270

83 Cohen P R, Johnston M, McGee D, Oviatt S, Pittman J, Smith I, Chen L, Clow J. QuickSet: multimodal interaction for distributed applications. In: Proceedings of the fifth ACM international conference on Multimedia. Seattle, Washington, USA, ACM, 1997: 31–40

84 Chatterjee I, Xiao R, Harrison C. Gaze + Gesture: Expressive, Precise and Targeted Free-Space Interactions. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. Seattle, Washington, USA, ACM, 2015: 131–138
DOI:10.1145/2818346.2820752

85 Velloso E, Turner J, Alexander J, Bulling A, Gellersen H. An Empirical Investigation of Gaze Selection in Mid-Air Gestural 3D Manipulation. In: Human-Computer Interaction – INTERACT 2015. Cham: Springer International Publishing, 2015, 315–330
DOI:10.1007/978-3-319-22668-2_25

86 Zhang F J, Dai G Z, Peng X L. A survey on human-computer interaction in virtual reality. Scientia Sinica (Informationis), 2016(12): 1711–1736
DOI:10.1360/N112016-00252

87 Wu L. Oviatt S L. Cohen P R. Multimodal integration-a statistical view. IEEE Transactions on Multimedia, 1999, 1(4): 334–341
DOI:10.1109/6046.807953

88 Sim K C. Speak-as-you-swipe (SAYS): a multimodal interface combining speech and gesture keyboard synchronously for continuous mobile text entry. In: Proceedings of the 14th ACM international conference on Multimodal interaction. Santa Monica, California, USA, ACM, 2012: 555–560
DOI:10.1145/2388676.2388793

89 Kopp S, Tepper P, Cassell J. Towards integrated microplanning of language and iconic gesture for multimodal output. In: Proceedings of the 6th International Conference on Multimodal Interfaces. New York, NY, ACM Press, 2004: 97–104

90 Ruppert G C S, Reis L O, Amorim P H J, de Moraes T F, da Silva J V L. Touchless gesture user interface for interactive image visualization in urological surgery. World Journal of Urology, 2012, 30(5): 687–691
DOI:10.1007/s00345-012-0879-0

91 Wachs J P, Stern H I, Edan Y, Gillam M, Handler J, Feied C, Smith M. A gesture-based tool for sterile browsing of radiology images. J Am Med Inform Assoc, 2008, 15(3): 321–323
DOI:10.1197/jamia. M2410

92 Keskin C, Balci K, Aran O, Sankur B, Akarun L. A Multimodal 3D Healthcare Communication System. In: 2007 3DTV Conference, 2007, 1–4
DOI:10.1109/3DTV.2007.4379488

93 Phelan I, Arden M, Garcia C, Roast C. Exploring virtual reality and prosthetic training. In: 2015 IEEE Virtual Reality (VR), 2015, 353–354
DOI:10.1109/VR.2015.7223441

94 Moustakas K, Nikolakis G, Tzovaras D, Strintzis M G. A geometry education haptic VR application based on a new virtual hand representation. In: IEEE Proceedings VR 2005 Virtual Reality, 2005, 249–252
DOI:10.1109/VR.2005.1492782

95 Vrellis I, Moutsioulis A, Mikropoulos T A. Primary School Students' Attitude towards Gesture Based Interaction: A Comparison between Microsoft Kinect and Mouse. In: 2014 IEEE 14th International Conference on Advanced Learning Technologies, 2014, 678–682
DOI:10.1109/ICALT.2014.199

96 Zhang X, Chen X, Li Y, Lantz V, Wang K, Yang J. A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 2011, 41(6):

1064–1076
DOI:10.1109/TSMCA.2011.2116004

97 Zhang F, Chu S, Pan R, Ji N, Xi L. Double hand-gesture interaction for walk-through in VR environment. In: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). Wuhan, IEEE, 2017, 539–544
DOI:10.1109/JCSSE.2015.7219818

98 Khundam C. First person movement control with palm normal and hand gesture interaction in virtual reality. In: 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE). Songkhla, IEEE, 2015, 325–330
DOI:10.1109/JCSSE.2015.7219818

99 Park H, Jeong S, Kim T, Youn D and Kim K. Visual representation of gesture interaction feedback in virtual reality games. In: International Symposium on Ubiquitous Virtual Reality. Nara, IEEE, 2017, 20–23
DOI:10.1109/ISUVR.2017.14

100 Latoschik M E: A gesture processing framework for multimodal interaction in virtual reality. In: Proceedings of the 1st international conference on Computer graphics, virtual reality and visualisation. Cape Town, ACM, 2001: 95–100
DOI:10.1145/513867.513888

101 Chun L M, Arshad H, Piumsomboon T, Billinghurst M. A combination of static and stroke gesture with speech for multimodal interaction in a virtual environment. In: 2015 International Conference on Electrical Engineering and Informatics (ICEEI). Denpasar, IEEE, 2015: 59–64
DOI:10.1109/ICEEI.2015.7352470