A Survey of Underlying Base Technologies for Virtual Experience

Jinseok Seo

Division of Digital Contents Technology, Dong-eui University, Busan, Korea

Abstract— *This paper is a survey of virtual experience provided by virtual reality systems. The author investigated the various underlying base technologies that enable virtual experience and the research results about the virtual experience systems using these technologies. The underlying base technologies that enable virtual experience is divided into recognition, generation, and expression technology. The author analyzed the three types of technologies in more detail and examined the state of the technical elements.*

Keywords— Virtual Reality, Virtual Experience.

I. INTRODUCTION

A virtual experience is a simulation of a real experience. Instead of experiencing and interacting with physical objects in a real environment, in virtual environments, we interact with specific media in place of physical objects in a real environment. It is the same that the external stimuli sensed by the sensory organs of the human body are transmitted to the brain, and the reaction initiated from the brain again is expressed through various body organs of the human body. On the other hand, from the viewpoint of the object of interaction, the means of sensing and expressing is different.

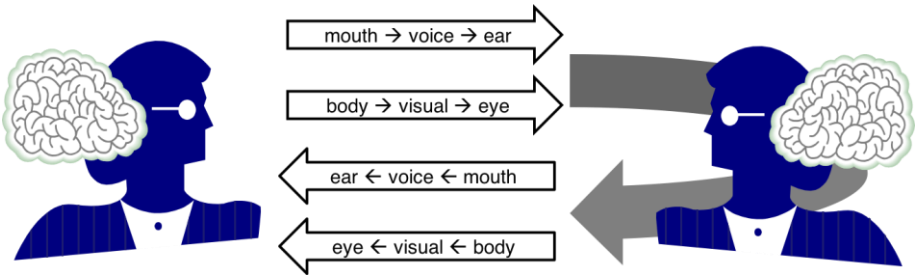

Fig. 1: Dialogue interaction between people and people

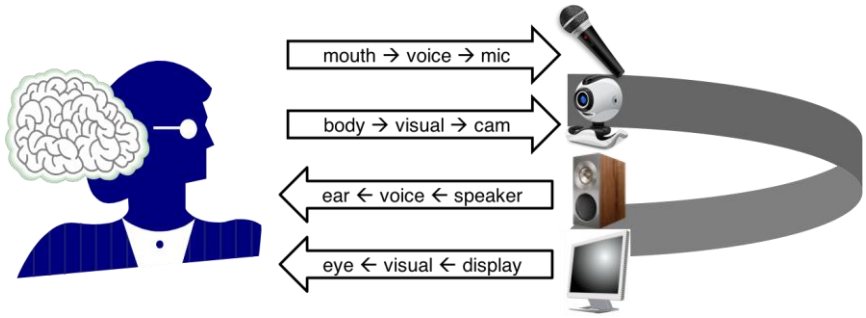

Fig. 2: Dialogue interaction between people and computers

For example, if a person is experiencing through an interaction method called "conversation," the actual environment and the virtual environment can be distinguished as shown in Fig. 1 and Fig. 2. In the actual environment [Fig. 1], the voice information generated through the mouth of the human body is transmitted to the ear of the opponent, and the visual information generated from the body (look, facial expression, behavior, etc.) is transmitted to the opponent's eyes. In the case of Fig. 2, it is assumed that the interaction target in the virtual environment is a general PC. Unlike humans, a person's voice information is acquired by a microphone and the visual information is acquired by a camera. Then, the acquired information processed by a PC's CPU. Response from PC, unlike humans, voice and visual information are represented by a speaker and a display monitor, respectively.

In order to provide the most ideal virtual experience, it is clear that the ability to recognize and express the object of virtual interaction must match that of human. To this end, scholars in the field of artificial intelligence have been studying for a long time to make computers and robots close to the human brain and in the field of virtual reality, efforts have been made to

establish a virtual environment that is maximally similar to reality. Though it does not yet offer the same level of virtual experience as real experience, decades of technological and academic advancement have made virtual experiences available in many areas.

II. UNDERLYING BASE TECHNOLOGIES FOR VIRTUAL EXPERIENCE

The underlying base technologies of virtual experience can be divided into three areas as recognition, generation, and expression [Fig. 3]. Recognition technology recognizes and digitizes voice and visual information expressed by humans, and generation technology generates new digital information in response to processing of recognition results. Finally, expression technology display digital information as voice or visual information that can be understood by humans. In Figure 3, although processing technology is an integral part of the system to provide a virtual experience, it is not an underlying technology for virtual experience only. The processing technology is an area of study with its own broader scope (e.g., artificial intelligence, signal processing, natural language processing Etc.). Therefore, it is excluded from this paper.

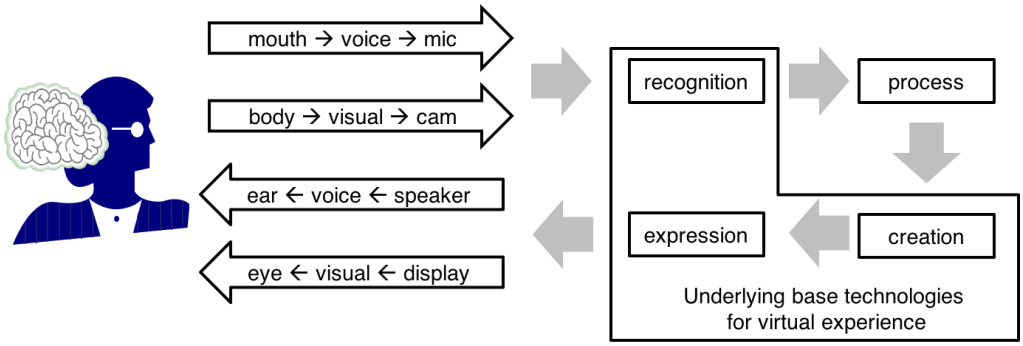

Fig. 3: Underlying Base technologies for virtual experience that can be divided into 3 areas

2.1 Recognition Technology

2.1.1 Sensors

Sensors are input devices that use various hardware technologies instead of conventional input devices, a keyboard and a mouse. In recent years, HW technology has developed at a high speed, and cheap sensors are being used in many parts of daily life. Typical examples include acceleration, gyroscope, RFID, touch, pressure, and temperature sensors.

In the case of gyro sensor, it is adopted in console game controllers and plays a role in driving the activation of experiential game contents. As gyro sensor is becoming smaller and lighter, it is widely used in mobile devices such as smartphones and tablet PCs. RFID has been used mainly for industrial purposes, but nowadays it is used as an auxiliary means for recognition of tangible objects in virtual experience-based contents.

Speech recognition is a technology that enables the most natural communication between a person and a computer. Speech recognition is classified into a speaker dependent, a speaker independent, and a speaker adaptive method. The speaker dependent method has a disadvantage in that only the voice of a specific speaker can be recognized although the accuracy is very high. In the speaker independent method, a learning process involving a lot of data is required in order to recognize an unspecified large number of voice. Speaker adaptation is a method that takes advantage of two methods. The speech recognition is utilized in various devices (computer operating system, navigation software, mobile phones, etc.), but it is fundamentally difficult to achieve 100% recognition rate. The cause may be pronunciation differences, hardware characteristics such as a microphone, and ambient noise.

2.1.2 Computer Vision

Computer vision originated in one area of artificial intelligence. The main purpose for obtaining information from an image requires the help of various theories such as pattern recognition, image processing, and graph theory. The computer vision in virtual experience is mainly used to recognize an object from real-time image and acquire position of the recognized object. Especially, it is an essential technology for making virtual objects in augmented reality contents.

Using multiple cameras, it is possible to track the exact location in 3D space. However, due to cost and installation space issues, the technology that uses only one camera has been used in recent years. In computer vision, as with voice recognition, there are many differences in the recognition rate depending on the hardware characteristics of the camera and the noise from ambient lighting. Studies are also actively conducted to automatically correct the noise.

2.1.3 Motion Capture

Motion capture obtains the movement of a specific object (or person) as numerical data. The obtained data is used as motion data for moving a virtual. Motion capture can be divided into mechanical, magnetic, and optical, depending on the type of the device used. Although the mechanical type has the advantage of obtaining accurate motion data, it has a disadvantage that it is necessary to install (or wear) a heavy and inconvenient mechanical device so that it is utilized only in a special field. Despite the disadvantages of the magnetic sensor that distortion is likely to occur in the acquired data and that many wires must be used, it is inexpensive and widely used. In recent years, the development of technology has made widespread use of inexpensive and accurate optical methods, in which a marker is attached to a moving object or major parts of a person and a plurality of cameras are used.

2.1.4 Compensation of Sensed Data

The above recognition sensors do not always obtain ideal data depending on the physical characteristics of the devices and the conditions of the environment. For example, in the case of computer vision, if there is a change in illumination, there will be differences in the recognition result. In the case of a motion capture device, it is often the case that, due to the limitations of the capability of the device, it is possible to obtain distorted (noise added) data that is different from the data that the actual person acted on. In such a case, a technique capable of compensating the original operation data to a form as close as possible is needed.

2.2 Generation Technology

2.2.1 3D Mesh Generation and Reconstruction

We use 3D mesh representation which can render in real time for 3D objects in virtual space. A mesh is a data structure composed of a set of polygons consisting of vertices, edges, and faces. The points of the curved surfaces of a given 3D object are sampled to determine the coordinates of vertices, the connection structure between the vertices is defined as edges, and a polygon composed of three or more edges linearly approximates the curved surfaces. In the past, modeling of direct 3D meshes by modeling software such as 3D studio MAX or Maya was used. Although this method has an advantage that it is easy to model a simple object, it has a disadvantage that it requires a lot of manpower and time to model an object having complicated detail.

In recent years, 3D modeling techniques based on reconstruction of 3D objects have been widely used, for example, by making a copy of a real object with a material such as gypsum and then capturing the shape with a 3D scanner. Such a 3D scanning method can easily generate a complicated and sophisticated object in a mesh form, but it is difficult to define the texture mapping immediately because the structure of the mesh obtained from the scanning process is more complicated than the complexity of the original object. In order to overcome these drawbacks, researches on mesh parameterization, re-meshing, and mesh modification have been actively conducted.

2.2.2 Mesh Simplification

Mesh simplification is a technique used to quickly render objects with complex mesh structures. Especially, since the mesh obtained through the 3D scanner is generally more complicated than necessary, mesh simplification is often used to effectively represent the object with a simpler mesh through the mesh simplification technique. In general, the mesh simplification is performed by repeatedly performing a vertex down sampling operation which reduces vertices one by one, which is considered unnecessary. A typical vertex down sampling operation is an edge decay operation. In this operation, the problem of which edge to collapse is the problem of finding an edge that keeps the original shape as full as possible, even if the edge collapses and the vertex decreases. This problem is usually solved by introducing the concept of the squared error

distance proposed by Garland and Hackbert. If mesh simplification is performed through their proposed method, it is advantageous to obtain a fast and high-quality mesh simplification result.

2.2.3 Rendering Acceleration (Real-Time Rendering)

In the past, rendering was done with a fixed rendering pipeline process in which the pre-defined process could not be changed. A representative feature of the fixed rendering pipeline is that the color value of one pixel is determined not from the characteristics of light, normal, material, etc. at the point corresponding to the pixel but from the color values of the vertices constituting the polygon. Therefore, in order to calculate the color value of each pixel more accurately, it was necessary to express an object with a larger number of polygons. In recent years, the development of graphics processing units(GPU) that complement the drawbacks of these fixed rendering pipelines has led to the widespread use of a programmable rendering pipeline that enables more effective real-time rendering.

The most important part of the programmable rendering pipeline is a shader. A shader is usually composed of a vertex shader and a pixel shader. The vertex shader has a function to output the data of the vertices, including position, texture coordinates, normal, etc., received as input, in a modified form through a shader program. The pixel shader has a function that allows the color values of each pixel constituting the screen to be calculated independently by a shader program. Since the color value of each pixel can be calculated using the data contained in the texture and the data from lighting, the effect of the pixel shader is generally determined according to how the texture is structured. For example, when the normal data of an object is stored in a texture and input to a pixel shader in the form of a normal map, a pong shading capable of rendering an object shiny can be implemented in real time.

2.2.4 Collision Detection

In order to construct a virtual environment close to the real world, a technique of simulating the collision and reaction of objects is required. It usually takes a lot of time to solve the problem of determining whether two objects have collided or not. For example, the following simple method can be considered to determine whether two objects represented by meshes have collided with each other or not. After two objects are represented by meshes A and B, it is checked whether a triangle of the mesh A collides with the triangles of the mesh B, and this process is repeated for all the triangles of the mesh A. To confirm that two objects did not collide, you should get a result that there is no collision in the collision check for all triangles. However, this operation takes a lot of time since it has to be performed as many times as the product of the number of triangles forming two objects. Therefore, there is a need for a technique that shows that there is no collision between two objects in real time, because a lot of computation time is required.

In order to check rapidly whether there is a collision, the following approximate method is used. The most common method is to use a bounding box or a circle. It is assumed that there is a box or circle surrounding the object, and the collision check between the boxes or circles surrounding the two objects is performed first before the collision check of the triangles of the two objects. If the object is far away, it can be easily judged that there is no collision between two objects. If a collision occurs with a bounding box or a circle, it is often the case that the collision between two objects is to be approximated or precisely determined depending on the application. In a virtual environment that does not require high precision, the collision of two objects is determined by a bounding box or by a bounding circle.

2.2.5 Simulation

Simulation is a technique to simulate the behavior of objects in virtual space. It requires theoretical knowledge on physics and mathematics. The goal is to make the motion of the object as similar to the objects in the real environment as possible. For example, in order to portray as realistically as possible the motion of a car being steered by a user, complex kinematics and dynamics must be mobilized, and in the case of a self-moving vehicle, techniques related to artificial intelligence agents are also required. Usually, virtual reality systems have used a lot of computing resources for visualization (rendering) rather than simulation, but recently, due to advances in computer graphics related theories and hardware advances such as GPUs, much research about simulation is under way.

2.2.6 Authoring Tools

Complex knowledge of various fields is required to construct a virtual environment. Typical examples include computer programming, 3D graphics theory, understanding of sensors and various input / output devices, and network theory. In order to build a virtual environment 10 years ago, many experts in various fields were required to spend a lot of time, but in recent years, various kinds of authoring tools have emerged due to the necessity of authoring tools. An example of a low-level tool is a class library in C++ language, and high-level tools such as game / virtual reality engine provide more integrated functionality. High-level tools include GUI-based authoring tools and scripting engines. In a near future, as the authoring tool technology becomes more advanced, it is expected that it will be possible to easily create a virtual experience environment without specialized knowledge.

2.3 Expression Technology

2.3.1 Stereoscopic Display

Although the display device of a computer can basically express only 2D images, we can mimic stereoscopic images by showing different images to two eyes. The stereoscopic image presentation technology can be divided into passive and active modes. The passive method is a technique of attaching different polarizing filters to two projectors for displaying the image for the left eye and the image for the right eye, and we see the stereoscopic image by wearing polarized glasses. This is the most popular and cheaply configurable way, but it has the disadvantage that it cannot be used for a long time due to ghost or flicker phenomenon. On the other hand, the active method uses a high-performance projector capable of displaying images at a high frequency and does not require a polarizing filter, but a user must use expensive glasses such as shutter glasses.

2.3.2 Immersive Display

An immersive display is a technique for providing a sense of realism to a user as if a user is directly in the virtual environment. There are two main types, such as head-mounted display (HMD) and head-mounted glasses. The HMD can be divided into stereotypes and monotypes depending on whether stereoscopic images are displayed or not. The stereotypes have two built-in LCD screens to display stereoscopic images, but they still require very precise optical technology to provide a wide field-of-view (FOV) and undistorted images, which are still expensive. As with mixed reality, there are special HMDs such as optical see-through HMD and video see-through HMD to display external images together. In the optical type, since the display device itself is made semi-transparent, the user has an advantage that the external image can be directly seen. On the other hand, the video type is a method of acquiring an image from a separate camera and synthesizing it with graphic contents. However, a parallax error can occur due to the difference of physical positions between the camera and the user's eyes.

Immersive environments using large displays mainly use projectors and screens. The display capabilities of the computers we use have limited resolution, which makes it difficult to create a large immersive environment. In the early days, a very expensive graphics workstation was used to display multiple video output devices, but in recent years, PCs and graphics accelerators have been developed to bundle multiple PCs (clustering) into an immersive display. In this case, each PC provides a different view, and synchronization between the PCs is the most important technology. Unlike the immersive environment using a large-screen display device, there are CAVE-like systems and curved displays as described below as a method of using specially arranged screens.

2.3.3 CAVE-Like Systems

As one form of immersive display devices, a screen is arranged in the form of a cube and the image is projected on the back side of the screen by a projector. The user enters the inside of the cube and experiences an immersive environment. This system is mainly composed of an active stereoscopic image system. CAVE-like systems are theoretically possible up to six-sided, but four-sided CAVE is common because of space constraints for projectors and mirrors placements.

2.3.4 Curved Display

This is similar to CAVE-like systems, but the screen is curved. A large number of projectors must be used to project on the curved surface, and very precise design is required to accurately project the image on each curved patch. Since the image projected on the curved surface is distorted, the image generation technique considering the curvature is very important.

III. SUMMARY

TABLE 1
UNDERLYING BASE TECHNOLOGIES FOR VIRTUAL EXPERIENCES

Category	Technology	Description
Recognition	Sensors	Recognizes various types of external signals
	Computer Vision	Extracts information from an image
	Motion Capture	Obtains the movement of an object as numerical data
	Data Compensation	Compensates the sensed data
Generation	Mesh Generation	Generates a data structure to represent a 3D object
	Mesh Simplification	Simplifies an complex mesh structure
	Rendering Acceleration	Enables real-time rendering of complex virtual environments
	Collision Detection	Checks whether one object is collided with another object
	Simulation	Simulates the behavior of objects in virtual space
	Authoring Tools	Provides a GUI-based authoring environment and a scripting engine
Expression	Stereoscopic Display	Mimics stereoscopic images by showing 2 different images to each eye
	Immersive Display	Provides a sense of realism to a user as if the user is directly in the VE
	CAVE-like System	Projects images on the back side of 4 or 6 screens in cube form
	Curved Display	Projects images from multiple projectors onto multiple curved screens

In this study, I investigated the underlying base technologies of virtual reality systems that provides virtual experience to users. Prior to conducting the survey, we analyzed the base technologies in three categories: recognition, generation, and expression, from the perspective that people and computers exchange information. Recognition technology recognizes and digitizes voice and visual information expressed by humans, and generation technology generates new digital information in response to processing of recognition results. Finally, expression technology display digital information as voice or visual information that can be understood by humans. Table 1 summarizes the detailed underlying base technologies included in each category.

REFERENCES

[1] Di Blas, N., Paolini, P., & Poggi, C. (2005). 3D Worlds for Edutainment: Educational, Relational and Organizational Principles, Proceedings of the 3rd International Conference on Pervasive Computing and Communications Workshops.

[2] Harada, Y., Nosu, K., & Okude, N. (1999). Interactive and Collaborative Learning Environment using 3D Virtual Realty Content, Multi-Screen Display and PCs, IEEE 8th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises.

[3] Ishii, H. (1998). Tangible Interface for Remote Collaboration and Communication.

[4] Kaufmann, H. & Schmalstieg, D. (2006). Designing Immersive Virtual Reality for Geometry Education, Proceedings of the IEEE Virtual Reality Conference.

[5] Lindinger, C., et al. (2006). Gulliver`s World: A case study on collaborative edutainment at the intersection of material and virtual worlds, Virtual Reality (2006) 10, pp. 109-118.

[6] Liu, T., Tan, T., & Chu, Y. (2007). 2D Barcode and Augmented Reality Supported English Learning System, 6th IEEE / ACIS International Conference on Computer and Information Science.

[7] Milgram, P., & Keshino, F. (1994). A taxonomy of mixed reality visual display. IEICE Transactions on Imformation and Systems, E77-D, 12, 1321-1329.

[8] Nishida, Y., Hiramoto, M., Kusunoki, F., & Mizoguchi, H. (2005). Learning by Dong: Space-Associate Language Learning Using a Sensorized Environment, Proceedings of IEEE International Conference on Intelligent Robots and Systems, pp. 1583-1588.

[9] Nitta, T., Fujita, K., & Kohno, S. (2000). An application of distributed virtual environment to foreign language education, 30th ASEE / IEEE Frontiers in Education.

[10] Rehm, M., et al. (2006). Location-Based Interaction with Children for Edutainment, LNAI 4021, pp. 197-200.

[11] Tolias, D. & Exadaktylos, G. (2007). Learning Through Exploration, Autonomy, Collaboration, and Simulation: The 'all-in-one' Virtual School of the Hellas Alive! Online, Language-Learning Platform, LNCS 4556, pp. 823-832.