Development of A Finger Pointing Interpretation and Speech Recognition System to Operate Virtual World

Takeshi Shiofuku , Norihiro Abe , Yoshihiro Tabuchi
Kyushu Institute of Technology
680-4 Kawadu, Iizuka, Fukuoka 820-8502, Japan
Tel : 81-948-29-7776
shiofuku@sein.mse.kyutech.ac.jp,

Hirokazu Taki
Wakayama University
930 Sakaedani, Wakayama-shi
Wakayama 680-8510, Japan
taki@sys.wakayama-u.ac.jp

Shoujie He
VuCOMP,
hesj@computer.org

Abstract:
In this research, the system constructed achieved the goal of real-time interaction with the virtual world through the interpretation of finger point gestures and speech recognition. Recently, Interaction with a computer is, however, mainly through a mouse and keyboard, which are not user-friendly enough for the majority. So, it is imperative to develop straightforward and easy-to-use user interface for computers. We made the system that was able to operate virtual objects with pointing action and voice without using an expensive, special device such as cyber gloves. The system was divided into voice recognition, pointing gesture recognition, and virtual world managing system as the processing time grew if these systems were executed with one PC. Data was shared by communicating data among them, and the decentralization allows the system to work in real time. The user is now able to specify a virtual object to be operated with pointing action, and will be able to operate it by voice.
Keywords: Pointing action, Speech recognition, Real time

1. INTRODUCTION

With the advancement of data processing and telecommunication technologies, more and more people are using the Internet in their daily life. So far, the Internet had been mainly used for collecting information and sending or receiving email messages. The widely deployed fiber optics based high-speed communication networks and the lower-price and high-performance computers make it possible to telecommunicate the data-intensive content such as voice, images, and video over the Internet. At the same time, VR technologies that used to be time-consuming in data rendering also had attained significant progress.

Interaction with a computer is, however, mainly through a mouse and keyboards, which are not user-friendly enough for the majority. People who are used to such devices could easily collect huge amount of information through the Internet. On the other hand, people who are not familiar with the usage of them could hardly benefit from the availability of the Internet technologies. In order to improve the unbalanced situation in terms of information collection, it is imperative to develop straightforward and easy-to-use user interface for computers.

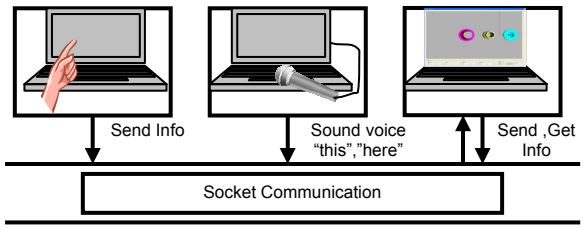


Figure 1: Concurrent processing among 3 PCs.

Then, in this research we developed pointing-recognition and speech-recognition system to operate virtual objects in virtual space. This system consists of three parts: In the vision part, user's pointing action is recognized based on image information obtained from cameras. The language processing part which

©ISAROB 2009

recognizes user's voice acquired with a mike. The display part shows a virtual world and allows a user to convey the requirement from a user to the system by using voice and action.

2. System Configuration

In this system, the application program is written in C++. The development environment is Microsoft visual C++.Net and we used the voice recognition soft that is called Dragon Naturally Speaking. Then, in order to recognize the gesture of pointing, we used 2 CCD cameras.

Table 1 : System configuration

Graphic library	OpenGL
Gesture recognition	2 CCD Cameras
Develop environment	Microsoft visual C++ . NET
Voice recognition	Dragon Naturally Speaking
Data communication	Socket communication(UDP)

This system does the parallel processing while communicating data each other among three systems managing pointing recognition, voice recognition, and a virtual space, respectively.

2.1 Research Environment

Figure 2 shows the research environment. The two cameras are set at the left side with the optical axes in parallel. It is supposed no fresh colored object to be put in the background.

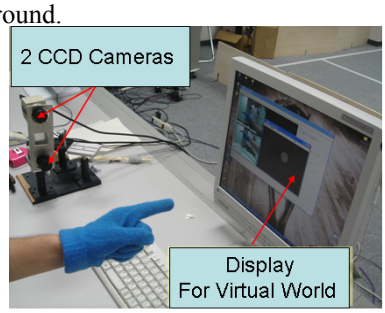


Fig2: Research Environment

3 Theory

3.1 UDP Communication

We use UDP communication protocol for the synchronization among the three PCs. In order to improve reliability, memory of each PC is always updated to the fresh data whenever the system writes data into memory.

3.1.1 Multicast

UDP can simultaneously transmit data to multiple destinations with the multicast option.

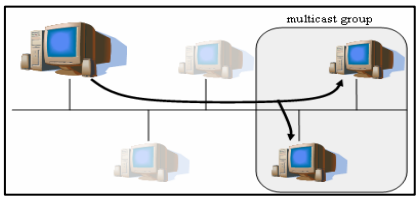


Figure 3: Multicast

In the multicast communication, data are transmitted to the PCs that participate in the multicast group (Figure 3).

- Reliability of data

UDP is a high-speed communication, but it is lack of reliability. The following approach has been taken to improve the reliability.

- Distinction of Information

If only a single port is available, information collides and the data loss will occur at the time when data are transmitted from VR, Voice, and Vision unit (Figure 4).

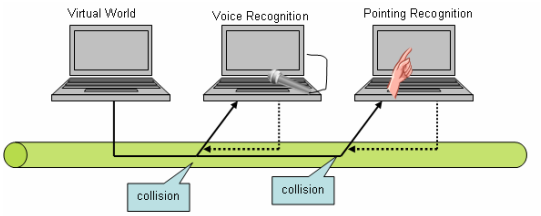


Figure 4: Collision of data

To avoid the collision, three ports are exclusively prepared for VR, Voice recognition system, and Pointing recognition system (Figure 5). This way prevents the data collision from happening.

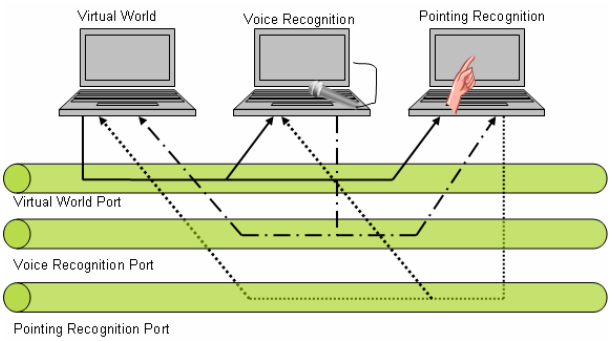


Figure 5: Each port

3.2 Pointing information acquisition

In order to detect the pointing direction, the system must extract flesh-colored regions from the image and assign a label to each region by removing noises. To calculate the vector of a forefinger, two characteristic points of the forefinger are detected.

3.2.1 HSV conversion

The original image data captured from the camera are in RGB color system. It is, however, difficult to set the range of the target colors in this color system.

This problem could be resolved by using the HSV color system. In the HSV system, the person can express the color sensuously. It is necessary to convert the image data from the RGB color system to the HSV color system. In the HSV color system, it is empirically known that a flesh-colored area is within the following ranges. Figure 3 shows the result with the flesh-colored region extracted.

Hue : -75 < H < 40 (Range : 0~360)
Saturation : 40 < S < 200 (Range : 0~255)
Varue : 0 < V < 255 (Range : 0~255)

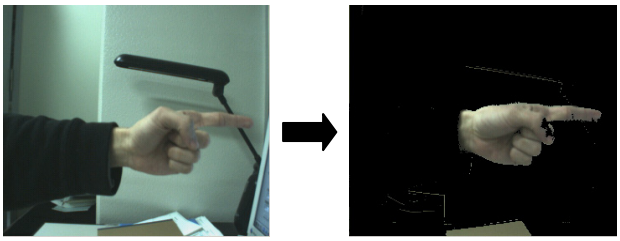


Fig6: The extracted flesh-colored image

3.2.2 Labeling

Labeling is a process of assigning a different label to each of the connected components. In this research, we used the run-length method in which the regions with a small area were removed.

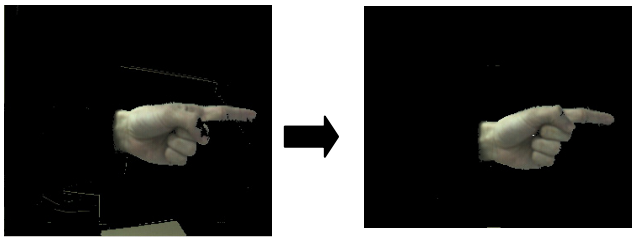


Fig7: The labeling

3.2.3 Characteristic Point Detection

The first characteristic point of a finger is a fingertip point. The other point is decided by comparing the amounts of the pixel forming the finger tip to back the hand.

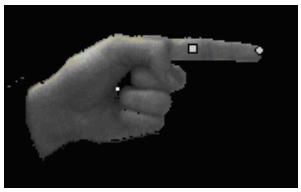


Fig8: Characteristic Point

The comparison equation is as follows. The index i is taken in a horizontal axis and x_dis[i] shows amount of pixels.

$$x\_dis[i] > x\_dis[i+5] * 1.5 \quad (1)$$

3.2.4 Stereo vision

Stereovision is a method of observing the same object from two different view points and measuring three dimension position of the object by the parallax of characteristic points. As shown in Figure 8, the three-dimensional coordinates of object P can be measured with the formulae (2), (3) and (4). Length in Figure 9 shows the distance between cameras, and f shows the focal length of the cameras.

$$X = Length\frac{(XL + XR)/2}{XL - XR} \quad (2)$$

$$Y = Length\frac{YL/2}{XL - XR} \quad (3)$$

$$Z = Length\frac{f}{XL - XR} \quad (4)$$

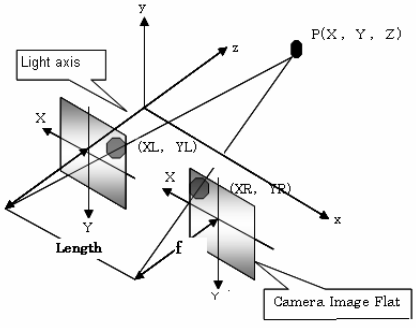


Fig9: Three-dimensional positional coordinates

3.3 Speech Recognition System

Speech recognition is achieved with a software package named Dragon Naturally Speaking. Dragon Naturally Speaking attains highly accurate voice recognition, and can translate the input voice into a sentence consisting of hiragana and Chinese characters. The speech recognition system retrieves the word from the set of words registered beforehand. The registered

©ISAROB 2009

word is limited only to a word set necessary for operating virtual objects.

For example, if the inputted sentence is "Move the paper to the position", the following words are detected: "Move", "Paper" and "Position" from the input voice.

3.4 Virtual World Operating System

The virtual space is constructed with OpenGL, a well-known graphics library. With this rendering technology, it is possible to construct a 3D world, where complex spatial relationships that are impossible to represent in a 2D world are easily handled. This system manipulates virtual objects according to the instruction recognized by the above-mentioned pointing recognition and speech recognition system. Concretely, this system determines which point within the display the user specified by referring to the coordinates acquired by the pointing action recognition system.

If the point is included in one of regions shown in. [Fig10], the system changes the current view point into the point where the user wants as shown in [Fig.11], but the point is included in the region of a virtual objects and if voice input includes a demonstrative pronoun, the system interprets that the user specified a virtual object to be operated.

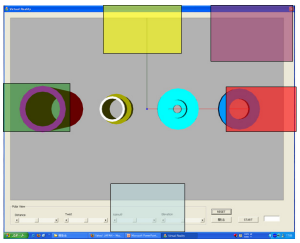
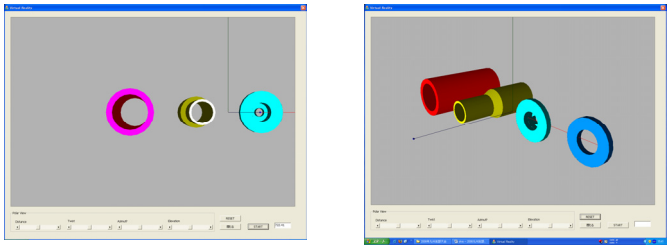


Fig10:Eyes Moved Area



(a)Eyes view movement to the left (b) Rotation around axis
Fig11: Eyes View Movement

4. Conclusions

The system constructed in this research achieves the goal of real-time interaction with the virtual world through the interpretation of finger point gestures and voice without using expensive tool such as a data glove. The translational/rotational movement of a viewpoint in the global coordinate system and manipulation of virtual objects in the local coordinate system are both performed in real-time. Since these operations make assembling/disassembling virtual object intuitive and easy, the high reality feeling as if real objects were manipulated is attained.

5. Future Work

Experimental results shows that the blur of finger caused by hand motion directly affect the accuracy in extracting feature points of a forefinger. In addition, to give much more real feeling, it is necessary to introduce collision detection between virtual objects. This is an issue in the future.

6. Acknowledgement

We greatly appreciate the Grant-in-Aid for Scientific Research(S) and (A).

7. REFERENCES

[1]Yoshiaki Shirai ,Masahiko Taniuchida "Pattern information processing" ohmsha,1998

[2]Yasuhiro Watanabe, Norihiro Abe, Kazuaki Tanaka, Hirokazu Taki, Tetsuya Yagi,Multimodal communication system allowing man and avatar to use voice and beck,3rd International Conference on Information Technology and Applications (ICITA'2005), pp.161-166

[3] Tatsuya Yoshikawa, Syunji Uchino, Norihiro Abe, Kazuaki Tanaka, Hirokazu Taki, Tetsuya Yagi, Shoujie He: Voice and gesture recognition system facilitating Communication between man and virtual agent, INVITE2006 International Conference on Parallel and Distributed Systems), Vol.2, pp.673-677, 2006

[4]Syunji Uchino, Norihiro Abe, Shoujie He, Hirokazu Taki, "Real-time Interactive Dialog System between Person and Virtual Agent", accepted, International Symposium on Artificial Life and Robotics (AROB 2007)