VR Interactive Dialog System with Verbal and Nonverbal Communication

Shunji UCHINO, Norihiro ABE, Yoshihiro TABUCHI

Kyushu Institute of Technology,
680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan
(Tel: 81-948-29-7776)
(Email: s_uchino@sein.mse.kyutech.ac.jp)

Hirokazu TAKI

Wakayama University,
930 Sakaedani, Wakayama-shi, Wakayama, 680-8510,
Japan

Shouji He

Eastman Kodak Company,
Plano, Texas,
USA

Abstract: In this research, a dialog environment between human and virtual agent has been constructed. With the commercial off-the-shelf VR technologies, special devices such as data glove have to be used for the interaction. But it is difficult for anyone to manipulate objects on one's own. If there is a helper who has direct access to objects in a virtual space, we may ask him. The question, however, is how to communicate with the helper. The basic idea is to utilize speech and gesture recognition systems. We have already reported the above-mentioned result. Though, only Avatar can move a virtual object in a current system. The user cannot freely manipulate virtual objects. So, as a new attempt, we constructed the communication channel between the virtual space and the real world so that the virtual object could be manipulated. And, to develop a new system, we extend the existing system to internet meeting system allowing users in different places to interact each other with voice and pointing action with a fore finger.

Keywords: Virtual Agent, Avatar, Dialog Environment, Internet Meeting System.

I. INTRODUCTION

Recently, toward a ubiquitous network society, products are developed that have some excellent functions based on Information Technologies. In future, it is expected that these products will be much more complex to provide multifunction. Nevertheless, main interface of computer such as mouse and keyboard will remain unchanged and it will make it difficult for elderly person to operate it. Further it will become a cause of digital divides. Also, it will be difficult even for experts in operating computer as instrument developers to use it when they grow older. So, developing new interface is expected which helps everyone to operate a computer easily. In this paper, a dialog environment between human and computer is proposed which unifies the verbal information using the voice and the non-verbal information using a gesture, and verified the validity of the system. It permits a user and a virtual human rendered in display to communicate each other using pointing action and utterance. The configuration of this system is shown in Fig.1.

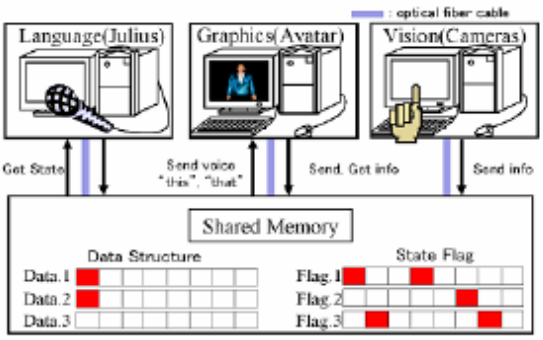

Fig.1. Concurrent processing among 3 PCs.

This system is divided into three parts: The language processing part recognizes user's voice picked from mike. In the vision part, user's action is recognized based on image information obtained from cameras. The display part shows a virtual human and makes him speak and act responding to requirement from a user. This system realizes actual dialog environment using both utterance and gesture by connecting these three parts with a high speed network called Scram Net+ which makes the time needed to send information among computers less than one millisecond.

II. SPACE FOR CONVERSATION

We have developed a system in which an avatar responds to a user as a salesman in a virtual fountain pen store. The user tells him the pen he wants to buy using utterance and a pointing action. The typical user's utterance with pointing action is like, "Please give this to me". The system recognizes the appropriate pen based on the verbal and non-verbal information. If the pen that the user wants cannot be located with the pointing action, the avatar must uniquely determine the pen with some questions. However, if he asks question many times, the user feels much annoyed. So some devices are necessary to reduce questions. We have already reported the method using a decision tree to have the avatar ask good question leading to unique identification of the pen.

III. INTERACTION FROM USER TO SYSTEM

Pointing action is required to synchronize with utterance such as 'What is this?' In this system, the spoken language processing and the gesture recognition are conducted concurrently to identify an object of pointing action. There is a possibility that the avatar may misunderstand user's instruction because pointing action involves ambiguity and thus pick up a wrong pen. In the case, the user has to immediately interrupt the action of the avatar and rectify wrong behavior.

1. Spoken language processing

Voice is analyzed with Julius for Windows version v3.3p4_jl2-1[1].

The technique to acquire the direction and coordinates of the user's pointing action is described in [2]. The user does the pointing action toward the screen as shown in Fig.2. The system extracts user's finger with a vertical stereovision system excelling in detecting a sidewise vector, and acquires the direction and coordinates of the user's fingertip and knuckle.

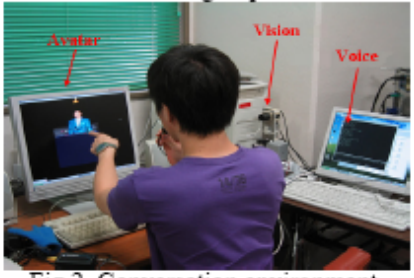

Fig.2. Conversation environment.

And we show the Stereo Vision System with two CCD cameras.

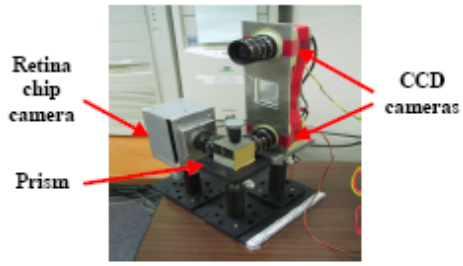

Fig.3. Stereo vision system.

2. Extraction of skin area

Generally color information data obtained from camera is RGB. It is difficult to extract a target object with peculiar colors from the environment illuminated with strong light. HSV color system helps an image processing system extract the target from such environment. Here, H of HSV, S and V means Hue, Saturation and Value of Brightness, respectively. The RGB color system is converted into the HSV color system. To extracts skin-colored area, it is necessary to set threshold values in terms of HSV value. Each of threshold values for extracting skin-colored area are shown below.

Hue: $0 < H < 40$ [Range: $0 < H < 360$]
Saturation: $0 < S < 140$ [Range: $0 < S < 255$] (1)
Value of Brightness: $0 < I < 255$ [Range: $0 < I < 255$]

Even if an effectual color values is given to extract skin-colored area, areas not corresponding to a hand are always extracted. The labelling process is useful to exclude all regions but the hand. This method is as follows.

1). Sequentially applying the label to each skin-colored region.
2). Extracting region of which area is the maximum value among them.
3). Making the image binary to remove noise except for the region with the maximum area.

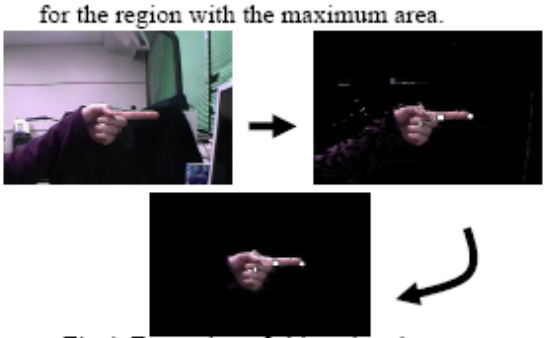

Fig.4. Extraction of skin colored area

IV. RESULT OF CONVERSATION

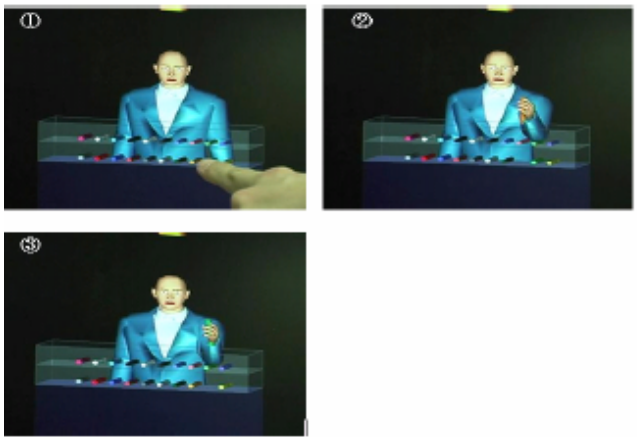

Fig.5. Flow of the conversation

Avatar: Which pen do you want to buy?
User: Please show me this one. (Pointing a gold pen)(1)
Avatar: Which color do you like?
User: Please give the gold one. (The avatar is going to reaching his hand to the pen.) (2)
User: Please give another one.
Avatar: Is this one? (Grasping another a pen) (3)

The above-mentioned content is a communication between User and Avatar. As a result, a user can feel actually as if he were talking to a real man. Thus, by constructing a natural conversation system with an avatar, it is possible for many users to use it as an easy interface instead of the mouse and the keyboard.

V. DIALOG SYSTEM WITH SOCKET COMMUNICATION

In the existing system, SCRAMNet was used to connect the PCs, which is very expensive. For this reason, we developed a system that enables the communication between a user and a virtual agent through the multi-port socket communication, instead of the SCRAMNet. This makes it possible that the regular PCs are used to achieve the same effects.

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.

1. Multicast

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

The way of the multicast works is that the data is transmitted to the PCs that participate in the multicast group.

Compared to TCP that transmits data to each PC sequentially, UDP had better performance in terms of the time needed for the data transmission.

2. Distinction of information

We made a two-way transmission port from each PC. If we use one single port, information collides and the data loss will occur at the time that data is transmitted from Avatar, Voice, and Vision unit (Fig.6).

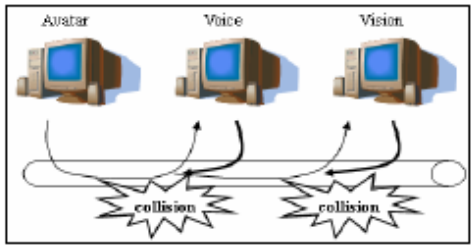

Fig.6. Collision of data

To avoid the collision, we prepared three ports, exclusively for Avatar, exclusively for Voice recognition system, and exclusively for Vision system (Fig.7). This way prevents the data collision from happening.

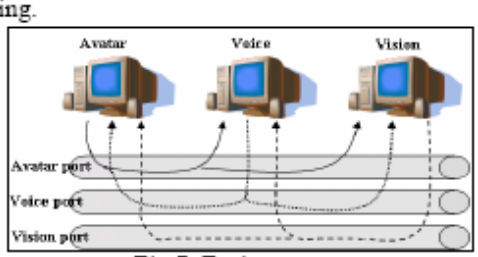

Fig.7. Each port

3. Combination of information

The next step is that we focused on the group of data stored in a memory location, and processed the data as follows. In the existing system, data is written in the integrated parts when we store data on the shared memory. The integrated parts are flags about the operation of Avatar, voice information, and so on. In this new approach, we transmitted the related group information when the content of one data is updated (Fig.8).

With this method, we are able to reduce the frequency of receiving data so that only the updated information will be received.

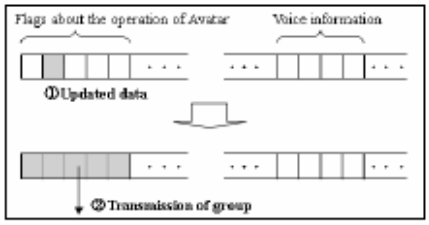

Fig.8. Transmission of group information

VI. INTERACTION BETWEEN VIRTUAL AND REAL SPACE

In the current system, only an avatar can move a virtual object. The human user is not allow to manipulate the virtual object freely. Consequently, we have been constructing the communication channel between the virtual space and the real world so that the virtual object could be directly manipulated by a user. The manipulation includes translation, rotation, and expansion and assembly operation of the object which permits a use to unit two objects with auxiliary line attached to them. Figure 9 shows the result of two parts assembled using two auxiliary line attached to them.

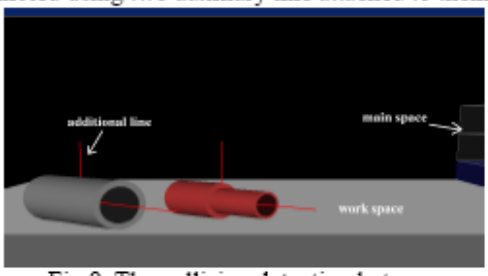

Fig.9. The collision detection between an object and additional line.

VII. INTERNET MEETING SYSTEM

A new internet meeting system among remote places is under construction that permits participants to communicate their intension with each other by using a pointing action and voice. We believe that this contributes to make the traditional net meeting much more vital and give users higher presence.

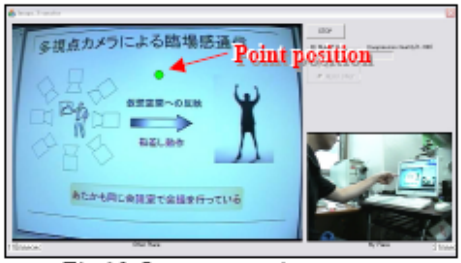

Fig.10. Internet meeting system with the pointing action

1. Communications system with SkypeAPI

We constructed the voice and video communications with Skype API [3] opened to the public, which allows a developer to operate Skype to transmit whatever data he wants in his any application. An advantage to exploit Skype API is that a developer is easily able to construct the system of the local LAN puss, to encipher the data and to distinguish the connected status with a companion.

In this system, the image and finger point data are easily sent or received among remote places. This will enable peoples in remote places perform verbal /nonverbal communication as if they were in a same room.

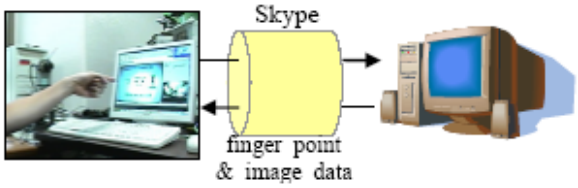

Fig.11. Content of communication data

VIII. CONCLUSION

We constructed the dialog environment in the real time between the user and the computer with the voice and gesture recognitions. Therefore, we developed the system that enables everyone to operate the computer easily. In the future, we advance the development of the above-mentioned internet meeting system with VR space.

ACKNOWLEDGEMENT

We greatly appreciate the aid of Ministry of Internal Affairs and Communications (MIC) and the Grant-in-Aid for Scientific Research (S) and (A).

REFERENCES

[1] Speech recognition software Julius,http://julius.sourceforge.jp/.
[2] T. Roppongi, "Pointing with retina chip camera and CCD camera and identification of the object"(in Japanese), Kyushu Institute of Technology, Master Thesis, 2003.
[3] SkypeAPI, https://developer.skype.com/Docs/ApiDoc
[4] T. Yoshikawa, S. Uchino, N. Abe, K. Tanaka, H. Taki, T. Yagi, S. He, "Voice and Gesture Recognition System Facilitating Communication between Man and Virtual agent", INVITE2006, Vol.2 pp.673-677, 2006.