Real-time Interactive Dialog System between Human and Virtual Agent

Shunji UCHINO, Norihiro ABE
Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
Email: s_uchino@sein.mse.kyutech.ac.jp

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

Shoujie He
Eastman Kodak Company,
Plano, Texas, USA

Abstract

In this research, a dialog environment between human and virtual agent has been constructed. With the actual VR technologies, special devices have to be used for the interaction with a virtual environment. This makes it extremely difficult for general users to manipulate an object in the virtual environment. In our daily life, when an object is out of our reach, we usually ask someone with direct access to the object to manipulate it on our behalf. If there is such a helper who has direct access to objects in a virtual space, we may do the similar thing. The question, however, is how to communicate with the helper, namely, the virtual agent. This paper presents a solution to the question. The basic idea is to utilize speech and gesture recognition systems and to integrate the verbal and non-verbal information. Experimental results have proved the effectiveness of the approach in terms of facilitating man-machine interaction and communication. The environment constructed in this research allows a user to communicate by talking and showing gestures to a personified agent in virtual environment. A user can use his/her finger to point at a virtual object and ask the agent to manipulate the virtual object.

Keywords: Virtual Agent, real time, voice and gesture recognition, interaction.

1 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 Figure 1.

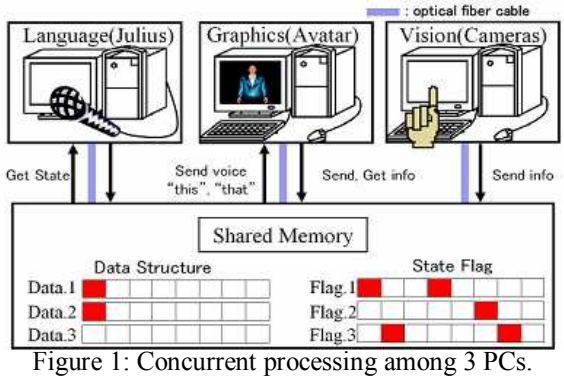

Figure 1: Concurrent processing among 3 PCs.

This system is divided into three parts: The language processing part which 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.

2 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.

3 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.

3.1 Spoken language processing

Voice is analyzed with Julius for Windows version v3.3p4_jl2-1. It is possible to recognize a given utterances at about 1.1-1.3 times of the utterance time. The starting and ending time of each demonstrative pronoun appearing in utterance detected with Julius are used to judge the temporal relation between utterance of a demonstrative pronoun and pointing action. This function successfully finds temporal relation even if utterance includes several demonstrative pronouns, and several pointing actions are given.

The user does the pointing action toward the screen. The system extracts user's finger with a vertical stereo vision system excelling in detecting of a sidewise vector, and acquires the direction and coordinates of the user's fingertip and knuckle.

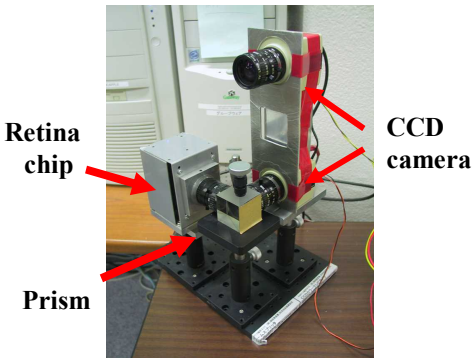

Figure 2: Combination of CCD and retina chip camera.

For the real-time processing, the system using only CCD color cameras is unsuitable. To reduce the cost of image processing, retina chip camera is used to narrow the image region to that including just a forefinger. The retina chip camera realizes a super-parallel image processing ("Moving object detection" and "Edge highlighting"…) with analog circuits embedded in each pixel. As shown in Figure 2, a ray passing through a prism is divided into two rays each of which is input to CCD and retina chip camera respectively. As the motion of a hand stops to make the object of pointing action clear, the still image of the hand is captured with a retina chip camera using images deference function. Images of two CCD cameras corresponding to the above retina chip image are examined to measure the direction and coordinates of the user's fingertip and knuckle.

3.2 Extraction of Skin Area

3.2.1 HSV Conversion

Generally color information data obtained from camera is RGB. It is difficult to exract 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 convered 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.

$$\begin{aligned} &\text{Hue:} && 0 < H < 40 \ [\text{Range: } 0 < H < 360] \\ &\text{Sarturation:} && 0 < S < 140 \ [\text{Range: } 0 < S < 255] \\ &\text{Value of Brightness:} && 0 < I < 255 \ [\text{Range: } 0 < I < 255] \end{aligned} \quad (1)$$

3.2.2 Labeling processing to remove noises

Even if an effectual color values is given to extract skin-colored area, areas not corresponding to a hand are always extracted. The labeling 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.

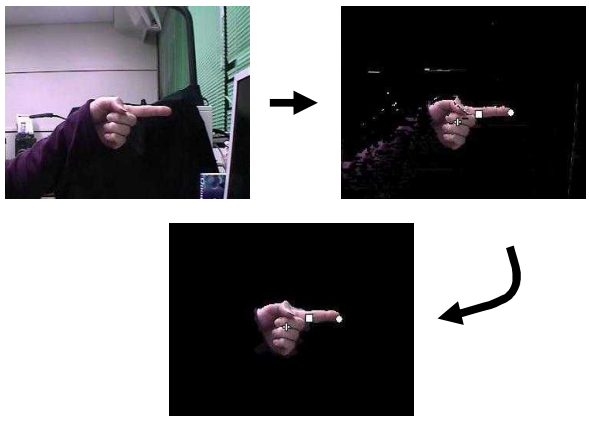

Figure 4: Extraction of skin colored area

3.3 Agreement of voice and action

The object in virtual space that the user specifies using pointing action can be determined based on the information through the spoken language processing. Even if a pronoun cannot be detected, when the pointing action is successfully found from the corresponding time span, it is possible to find the pronoun using the verbal-nonverbal relation described above and vice versa.

4 Result of conversation

Consider the case a user talks with an avatar in virtual space including 30 fountain pens in a showcase. Every fountain pen includes 6 attributes such as Size, Length, Color, Weight, Cost, and Product buy. It is possible to construct a decision tree and use the tree to deicide which pen a user wants to buy. But the method does not allow the user to specify freely his favorite attributes but forces him to answer whether he likes or dislikes a peculiar attribute which the decision tree algorism selects.

Instead, our system permits him to select his favorite pen by pointing action with the utterance such as "Would you show me this pen?" or "Would you take this and that pen here?" The first utterance must include just one pronoun and one favorite pen must be specified with pointing action. On the contrary, the second one include two pointing action but the discrimination of them is easy although it is not easy to find two pens if two pointing action specifies neighboring regions.

In both cases, the avatar has to find a set of candidate pens. For the first case, he can select two ways to determine what a user wants to buy. The first way is to ask if the user likes the one pointed or not with pointing one of candidates. This method is adequate if the number of the candidates is a few, otherwise the avatar may have to enumerate all of them. The second one is to make a decision tree from the set of candidates and then to ask the user which values his favorite pen has in terms of the attribute designated on the current node tracing the tree downward from the root. In this case, the number of inquiries to be asked will be smallest.

Here, the third one is proposed. Let us see an example. The flow of the conversation is shown in Figure 5. It is to use interrupt mechanism. Consider the following dialog in which a user specifies his favorite pen by pointing action (1). At this time, as 6 pens are selected as a candidate because of the ambiguity of pointing action, he asks the color to qualify a favorite pen. As a result, two red pens are left. Though he cannot decide which one is specified, instead of asking additional question he is unintentionally or intestinally going to grasp one of hem (2). Instantaneously, the user interrupts his action to inform him of negation. This makes him under-stand what the user wants. For confirming his favorite pen, he grasps the pen (3).

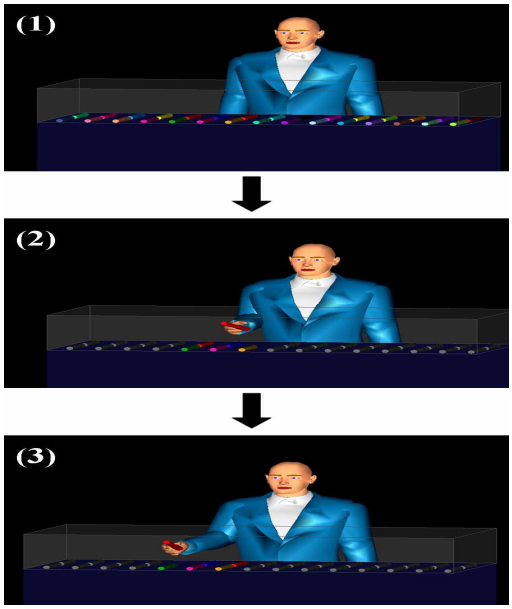

Figure 5: Flow of the conversation.

- Avatar: Which pen do you want to buy?
- User: Please show me this one. (Pointing a red pen) (1)
- Avatar: Which color do you like?
- User: Please give the red one. (The avatar is going to reaching his hand to one of two red pens.) (2)
- User: No, it is wrong.
- Avatar: Is this one? (Grasping another red pen) (3)
- User: Yes.
- Avatar: Thank you for purchasing this pen.

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.

5 Dialog system with socket communication

In the existing system, SCRAMNet+ was used to connect the PCs, which is costly. For this reason, we developed a system that enables the communication between a user and a virtual agent through the 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.

5.1 Multicast

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

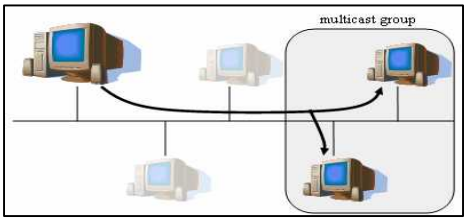

Figure 6: Multicast

The way of the multicast works is that the data is transmitted to the PCs that participate in the multicast group (Figure 6).

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

5.2 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.

5.2.1 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 (Figure 7).

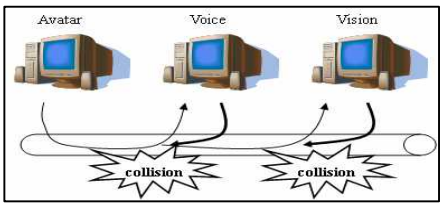

Figure 7: Collision of data

To avoid the collision, we prepared three ports, exclusively for Avatar, exclusively for Voice recognition system, and exclusively for Vision system (Figure 8). This way prevents the data collision from happening.

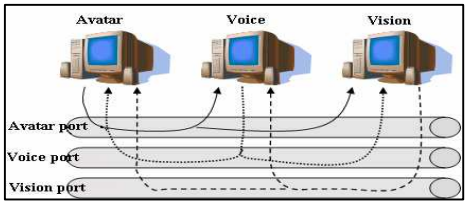

Figure 8: Each port

6 Conclusion

In this dialog environment, we introduced a virtual agent named avatar. So, 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. And, we completed a system that enables the communication through the socket communication, instead of the SCRAMNet which is costly. This result was able to be executed by equal ability with SCRAMNet. This made it possible that the regular PCs are used to achieve the same effects. However, in a current system, the human user cannot freely manipulate the virtual object. Therefore, we should construct the communication channel between the virtual space and the real world so that the virtual object could be manipulated. The manipulation includes translocation, rotation, and expansion and contraction of the object. This is an issue in the future.

Acknowledgements

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

References

[1] Ministry of Economy, Trade and Industry, "Society (e-Life strategy society) basic strategy report concerning strategy of making to market of information appliances - e-Life initiative -", 2003.

[2] N. Abe, and Tsuji, “A consulting system which detects and undoes erroneous operations by novices”, Proc. of SPIE, pp.352-358, 1986.

[3] N. Abe, T. Amano, K. Tanaka, J.Y.Zheng, S. He, and Taki, “A Training System for Detecting Novice's Erroneous Operation in Repairing Virtual Ma-chines“, ICAT, pp.224-229, 1997.

[4] N. Abe, J.Y.Zheng, K. Tanaka, and Taki, “A training System using Virtual Machines for Teaching Assembling/Disassembling Operations to Novices“, Inter-national Conference on System, Man and Cybernetics, pp.2096-2101, 1996.

[5] K. Tanaka, T. Ozaki, N. Abe, H. Taki, “Verbal/Nonverbal communication between Man and Avatar in Virtual Mechanical Assembly Training System”, Proc. VARI, pp.202-207, 2002.

[6] T. Ozaki, “Verbal and Non-Verbal Comunication Using Virtual Reality“, Kyusyu Institute of Technology, Master Thesis, 2001.

[7] T. Yagi, “Silicon visual area”, Japanese Neural Network Society, Vol.8, No.2, pp.65-69, 2001