Active Agent Oriented Multimodal Interface System

Osamu HASEGAWA; Katsunobu ITOU, Takio KURITA, Satoru HAYAMIZU,
Kazuyo TANAKA, Kazuhiko YAMAMOTO, and Nobuyuki OTSU

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki, 305 JAPAN
Phone : +81-298-58-5944, Fax : +81-298-58-5949

Abstract

This paper presents a prototype of an interface system with an active human-like agent. In usual human communication, non-verbal expressions play important roles. They convey emotional information and control timing of interaction as well. This project attempts to introduce multi modality into computer-human interaction. Our human-like agent with its realistic facial expressions identifies the user by sight and interacts actively and individually to each user in spoken language. That is, the agent sees human and visually recognizes who is the person, keeps eye-contacts in its facial display with human, starts spoken language interaction by talking to human first.
Key words : AI application, Multimodal Interface, Autonomous Agent, Spoken Dialogue, Visual Recognition, Facial Display

1 Introduction

In normal human communication, face-to-face communication in particular, humans activate many communication modes/channels in parallel and exchange verbal and non-verbal signals. Sight and hearing are the examples of such modes. As a result, a message conveyed by such modes can be reinforced each other, so that communication becomes highly flexible and meaningful.

On the other hand, numerous researchers have studied and proposed a variety of intelligent agents [6]. It can be said that an essential common to all such research is to develop agents which engage and help all types of end users. Therefore, agents should have the capacity for user-friendly and easy-to-use interaction with users.

* hasegawa@etl.go.jp

Recently, in view of these facts, research into multi modal interaction has become popular. Such projects attempt to understand the characteristics and utilization of human modes, and to introduce knowledge regarding them into human-computer interaction. However, these aspects have, until now, remained open problems.

We have been working on these problems. For example, we have collected and analyzed data on human behavior during interactions with a simulated spoken dialogue system [5]. We are on the way to the development of a mathematical model which describes the activation, integration, recognition and learning process of multi modal information. However, it is necessary to evaluate the mathematical model developed. For mathematical modeling and its evaluation, a multi modal interaction system should be developed and examined by all types of end users.

In this paper, we describe an active agent oriented multimodal interface system with image and speech recognition/synthesis functions as the first prototype of the research project.

The prototype displays a moving human-like agent with realistic facial expressions to promote smooth interaction with users (see Section 3). The agent can identify the user on sight and provide active interaction to users in spoken language. That is, the agent can execute such tasks as follows.

1. The agent starts spoken language interaction by talking to human first.
2. The agent sees human and visually recognizes who is the person.
3. The agent keeps eye-contacts in its facial display with human.

As a result, the agent can provide individual interaction with each user. In other words, the agent can respond differently to each individual user. These

are achieved by the integration of image and speech recognition/synthesis technologies.

2 Related Works

The idea of introducing a human-like agent into human-computer interaction was proposed in the mid 1980's by Alan Kay. However, at that time, it was hard to develop agents which could provide "ordinary" interaction modes for users because the computing power and costs in the mid 1980's were not prohibitive.

John Sculley, for example, proposed a human-like agent called Phil in the " Knowledge Navigator" in 1987, but it was demonstrated only in the concept video, and the utilization of visual functions was not considered.

Nevertheless, as the 1980's move into the 1990's, realistic agents with some interaction modes have begun to appear. In the following, some representative researches which utilizes human-like agents are refered to.

Takebayashi et al. have developed a spoken dialogue system (SDS) with a cartoon like but moving facial display (agent) [10], As their system does not have visual functions, it tries to find a user using a special switch in a floor mat.

Nagao et al. have developed a SDS with a human-like agent which joins human-human conversations and presents beneficial information for users [9, 11]. The agent is texture mapped with a real human texture, and the appearance is realistic. However, as this system does not have visual functions, the agent determines the presence of the user(s) from his/her voice.

Maes et al. are developing human-like agents which assist users with daily computer-based tasks [8]. In their framework, multi-agent collaboration is discussed based on learning agents. However, only simple caricatures are used to convey the state of the process (agent) to users. As for input from users, only the standard devices (keyboards and others) are used. Therefore, the channels between human and computer are not sufficient for natural interaction with users.

The Apple Newton with its *agent software* and General Magic's *messaging agents* are (will be) marketed as commercial products which employ human-like agents, but neither visual or speech recognition functions are supported.

The prototype system proposed in this paper can provide multi interaction modes (sight, hearing, speaking, and facial expressions) which are essential for intimate communication between humans and computers.

The facial image synthesis technique applied here is also used in videophone and video conference communication. However, our facial display is an autonomous agent, not a duplication of the user's face.

(This paper does not deny the utilization of keyboards and pointing devices. The important point is whether the interaction system can give users opportunities to select acceptable modes or devices, or not.)

3 Why Human-like Agent?

In a previous research project, an experiment was carried out to collect data on human behavior in interaction with computers [5]. In this experiment, forty subjects were requested to speak with a simulated spoken dialogue system. (The system used in this experiment did not display any visual agents.)

After the experiment, information was obtained from the subjects by questionnaires. The following results were obtained:

1. Seven subjects voluntarily requested to display facial symbols to speak with.
2. If the subjects feel that the system behaves like a human, they feel intimacy towards the system.

Based on this information, it can be said that realistic human-like agent promotes human-computer interaction. As a result, it was decided to apply a realistic and human-like agent as an interface surface between humans and the computer.

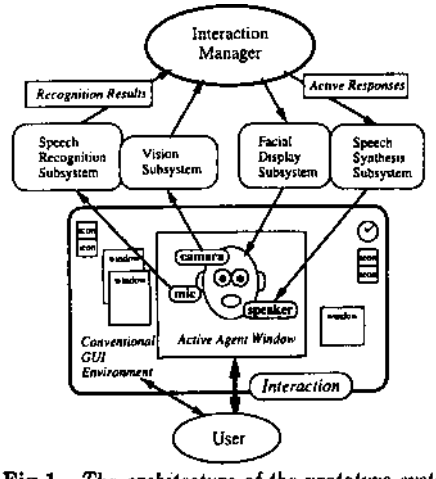

Fig.1 *The architecture of the prototype system.*

Fig.2 Facial texture and fitted SD model.

4 Prototype System

4.1 System Architecture

The developed system consists of three Work Stations for real time image and speech processing, one auto-focus CCD camera and one microphone. These are all standard hardware and equipment.

Figure 1 illustrates an outline of the system architecture. It consists of the following four sub-systems and a interaction manager.

1. A facial display sub-system that generates three-dimensional facial images.
2. A vision sub-system that recognizes and distinguishes users' facial images.
3. A speech recognition sub-system, that recognizes speaker- independent continuous speech.
4. A speech synthesis sub-system that generates voice output.
5. A interaction manager that controls inputs and outputs of the sub-systems.

The details of these sub-systems are described in the following subsections.

4.2 Facial Display Sub-system

The face of the agent is composed of approximately 500 polygons and is modeled three-dimensionally [2].

The appearance of the face is rendered by the texture mapping technique which is commonly used in computer graphics. The facial texture employed in our system is taken from a photograph of a young man. Figure 2 illustrates a 3-D facial model fitted onto the texture used in the system.

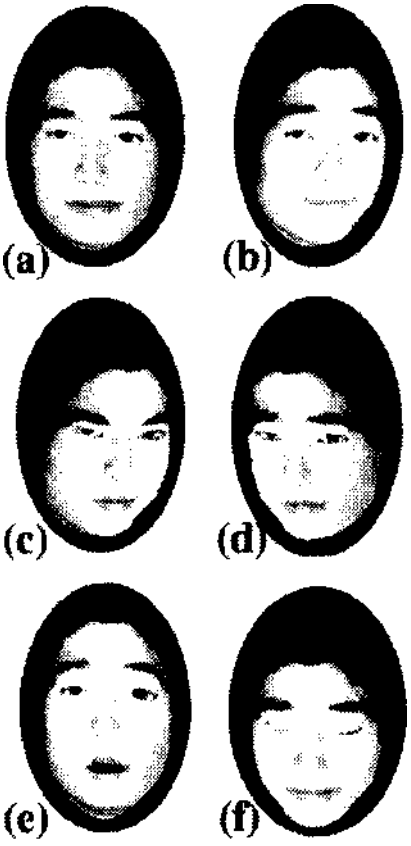

Fig.3 Samples of synthesized communicative facial displays (the agent), (a)neutral, (b)happiness, (c)anger, (d)sadness, (e)surprise, (f)sleep.

Facial displays are synthesized by local deformations and rotations of the polygons. Currently, eyebrows, eyeballs, eyelids, mouth and head orientation of the facial model are controllable. As a result, the action units in Facial Action Coding Units (FACS) [1] are available on the system.

Moreover, the system can control both action degrees and action speeds of each facial part indepen-

dently to provide realistic moving facial displays. Figure 3 shows samples of synthesized communicative facial displays (the agent). These are neutral, happiness, anger, sadness, surprise, and sleep.

It is common knowledge that "eye-contact" plays a vital role in human communication. For this reason, eye-contact was introduced into the prototype system, so that the agent may become friendly with users.

The current system controls the agent's eyes continuously so as to look at users during the interaction. Figure 4 shows a comparison between facial displays with and without eye-contact.

The number of parameters for the facial display is 24 for the current system. The base performance of the facial display sub-system is approximately 13 frames per second on SGI indigo2.

Currently, the system in-corporates thirteen parameter-sets (command sequences) for moving facial displays, considering the correspondence with tasks of the system (see Subsection 5.2).

The quality of the texture-mapped facial images is high, making them much more realistic than standard rendered images or animations.

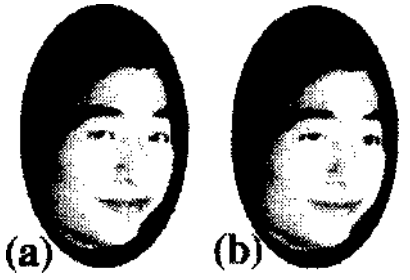

Fig.4 *Comparison between facial displays (a) with eye-contact and (b) vnthout it.*

4.3 Vision Sub-system

The purpose of the vision sub-system is not only to detect the presence of a user, but also to identify a facial image as a specific person.

User's facial images are taken from a standard video camera connected to the system. The camera is set beside the monitor, which displays the agent, so as to face the user.

The algorithm implemented is based on [Kurita et al.][7] so that real-time and robust processing will be available on the prototype system. The method employs higher order local autocorrelation features as primitive features at the first stage of feature extraction. Those features are then linearly combined using linear Discriminant Analysis or Multiple Regression Analysis to identify the user.

Thus, a moving face in input image sequence is recognized and identified in real-time. The background of the input images does not need to be constrained or segmentation free.

The recognition rates of the prototype system were approximately 98% for identification of 116 persons. The recognition speed was about 5 frames per second on a SUN sparcstation 10. This appears to be satisfactory for a human-computer interactive system.

4.4 Speech Recognition Sub-system

User's utterances are obtained via a microphone set in front of the monitor displaying the agent. Speaker-independent and continuous speech recognition is implemented in the prototype system. The speech recognition sub-system recognizes a (short) sentences one after another. The algorithm is based on [Itou et al.][4].

The recognition rate was approximately 84.2% spontaneous speech of for 40 subjects (183 utterances), but the experiments were carried out before the sub-system was integrated with the other sub-systems. Calculation time required for speech recognition ranges from 1-2 sec after the end of each user's utterance.

Currently, the sub-system can deal with a vocabulary of approximately 100 words and reject utterances that have low likelihood scores.

4.5 Speech Synthesis Sub-system

The speech of the agent is synthesized by the speech synthesis sub-system in a male voice. The sub-system synthesizes a speech consisting of one or more sentences at a time. However, it cannot control output timing precisely.

4.6 Interaction Manager

Currently, the whole system is controlled by the interaction manager.

The interaction manager receives messages (recognition results) both from the vision and speech recognition sub-systems in parallel. In order to obtain a high level of accuracy, it examines the order of the received messages and discards inadequate ones by following the state of the dialogue.

Then it analyzes the massages received and generates control commands for the facial display and speech synthesis sub-systems.

Followings are available basic tasks controlled by the interaction manager on the current system. In interaction, these tasks are executed in parallel. All tasks are executed with moving facial displays of the agent.

1. The agent finds a facial image, identifies it with a specific user and speaks to him/her actively.
2. The agent answers questions concerning the date and the time by speech.
3. The agent gives notice of incoming E-mail messages by speech.
4. The agent sets and explains the user's time schedule.
5. The agent records/replays messages from/to a specific user.
6. The agent sleeps between tasks.

5 Example Interaction with the Prototype System

5.1 Preliminary Arrangements

The experimental interactions are executed in an ordinary office environment.

Before interactions, the agent requires a learning process of users' facial images and a background. As described in Subsection 4.3, the background need not be constrained. In the learning process, approximately 50 images taken from a video camera are required for every user and the background. This process completed in a few minutes.

As for speech recognition, no preliminary arrangements are necessary.

5.2 Example of Interaction

This section describes an example interaction between two users and the prototype system. (The original dialogue is in Japanese.) In the following, A, Ul and U2 denote the agent, User 1 and User 2, respectively.

(Ul sits in front of the system
and looks at the monitor displaying the agent. The agent finds and identifies a facial image with U1 and speaks to him actively.)

A : Hello, Mr.U1.
U1: Hello.
What time is it now?
A : It is 14:30.
U1: Do I have any new mail?
A : No, you don't.
U1: OK, thanks

(U1 begins his daily computer-based tasks.)

A : New mail has arrived.
U1: Well..(U1 reads the mail.)
Now I 'm going out
A : What time will you come back?
U1: I'll be back at four.
A : I understand.
U1: Bye.
A : Good bye.

.....

(U2 comes and sits in front of the system and looks at the agent. The agent calls his name actively.)

A : Hello, Mr.U2.
U2: Hello.
Where is Mr.U1?
A : Mr.U1 will come back at four o'clock.
U2: I see. Thank you. Bye-bye.
A : Good bye.

Fig.5 *A user and the prototype system.*

5.3 Discussions

In the above dialogue, it is noted that the agent speaks to users *actively* and *individually* by making use of its "eyesight". It is one of the achieved tasks that users can leave (short) messages for a specific person with the agent.

However, the intelligence of the current agent is still limited. For example, should a recognition module make a mistake, the current agent cannot detect/recognize such mistakes. This remains as one of the problems to be solved. Figure 5 illustrates a user and the prototype system. The agent is displayed at the center of the monitor.

6 Conclusion and Future Tasks

We described the architecture and functions of our first prototype of an active agent oriented multi-modal interface, and discussed advantages and difficulties involved in such system. Through experimental interactions, we found that the activeness of the agent is effective not only in human-computer interaction but also in human-human communication via the agent.

We are planning to implement the following extended functions which enables the current agent more flexible and user-friendly. 1) Hands will be added to the agent. In human-to-human communication, gestures with hands play important roles as an additional mode which conveys non-verbal messages. 2) In image recognition, the algorithm will be improved to identify each user individually in the presence of plural users [3]. 3) In speech recognition, the number of words which can be dealt with will be increased.

We also plan to carry out experiments with subjects on the prototype system for the evalumation and improvement of the mathematical (interaction) models which is under development.

Active agents will be novel type of intelligence for the future information originated society.

Acknowledgments

This research project has been carried out as part of the Real World Computing (RWC) Program. The authors would like to thank those concerned. We also would like to extend our thanks to Prof. Hiroshi Harashima (Univ.of Tokyo) and his group for granting permission to use their 3D facial model on our prototype system.

References

[1] Ekman P. and Friesen W.V.:Facial Action Coding System, Palo Alto, CA, Consulting Psychology Press. 1978

[2] Hasegawa 0., Lee C.W., Wongwarawipat W., and Ishizuka M.: "Real- time Interactive System Between Finger Signs and Synthesized Human Facial Images Employing a Transputer-Based Parallel Computer', in T.L.Kunii Ed. Visual Computing. Springer Verlag. pp.77-94. 1992

[3] Hasegawa.0., Yokosawa K. and Ishizuka M. : "Real-time parallel and cooperative recognition of facial images for an interactive visual human interface" . Proc . of 12th ICPR, Jerusalem, Vol. 3, pp.384-387. 1994

[4] Itou K., Hayamizu S., Tanaka K., Tanaka H.:[1] System design, data collection and evaluation of a speech dialogue system", IEICE Trans, INF.&SYST., Vol.E76-D. No.I, pp.121-127, 1993

[5] Itou K et al.:'Collecting and Analyzing Non-verbal Elements for Maintenance of Dialog Using a Wizard of Oz Simulation', Proc.Int'l Conf. on Spoken Language Processing, pp.907-910, (Si 7-10.1 - S17-10.4), 1994

[6] Kay A.:"Computer Software", In Sci. America, 251, 3, pp.191- 207. 1984

[7) Kurita T., Otsu N. and Sato T. :"A Face Recognition Method Using Higher Order Local Autocorrelation And Multivariate Analysis", Proc. of IIthICPR, The Hague. Vol . II, pp.213-216, 1992

[8] Maes P.: "Agents that Reduce Work and Information Overload" In COMMUNICATION OF THE ACM. Vol.37, N0.7, pp.31-40. 1994

[9] Nagao K. and Takeuchi A.:"Social Interaction: Multimodal Conversation with Social Agents" Proc 12th AAAI 1992

[10] Takebayashi Y., Nagata Y. and Kanazawa H.: "Noisy spontaneous speech understanding using noise immunity keyword spotting with adaptive speech response cancellation" Proc IEEE pp.11115-11119, 1993

[11] Takeuchi A., and Nagao K.:"Communicative Facial Displays as a New Conversational Modality" in Proc INTERTECH ' 93 ACM Press pp.187-193, 1993