A Visual Software Agent: An Internet-Based Interface Agent with Rocking Realistic Face and Speech Dialog Function

Hiroshi Dohi **Mitsuru Ishizuka**
Dept. of Information and Commun. Eng.,
Faculty of Eng., University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113, JAPAN
E-mail: dohi@miv.t.u-tokyo.ac.jp

Abstract

This paper describes the Visual Software Agent (VSA), an Internet–based interface agent with rocking realistic face and speech dialog function. The VSA is connected with the WWW/Mosaic. Therefore, we can describe information for the VSA by Hyper-Text Markup Language (HTML), a widely used language for hyper-media in the WWW. A sub agent autonomously gathers information like weather forecast and news topics into a local database. It accesses periodically specific WWW servers, then picks necessary information out and converts it into a suitable format for the VSA. A user can also navigate on the Internet by speech dialog with the VSA, in addition to current mouse interface in the Mosaic. Both the speech dialog operation through the VSA and the mouse operation have the same priority. The user can choice the suitable operation method in any time in accordance with the various situations. It is useful for persons unfamiliar with a computer and physically handicapped persons, and for situations that the mouse interface is not suitable, for example, in a system using a wall-type room-wide display.

Introduction

As Internet-based information systems are rapidly popularized, software agents assisting users become important(Etzioni & Weld 1994; 1995). The word "*software agent*" generally means a computer software that performs intelligent tasks autonomously like human agents. In most cases, these software agents have no figure and no face. For example, Maes's Maxims system(Maes 1994) displays a simple caricature; however, this caricature itself is not an *agent*, but a simple indicator which conveys the state of the software to the user.

It is especially called "interface software agent" when a software agent takes care of interface tasks between a human and a computer(Laurel 1990). When an agent has a speech dialog function in addition to autonomous behaviors, its figure plays an important role.

Talkman(Takeuchi & Nagao 1993; Nagao & Takeuchi 1994) is an anthropomorphous agent system with speech dialog. It has a realistic 3-D facial image, and changes its facial expression according to the internal state of emotion. It has proved quantitatively that interfaces with facial displays reduce the mental barrier between users and computers.

It may be reluctant psychologically for a user to mutter to a simple physical box of computer. We feel uneasy whenever we speak to without a communication partner. An anthropomorphic agent with a realistic face embodies an environment close to "face–to–face" communication and "eye contact" between a human and a computer. It mitigates the user's unpleasantness when a speech dialog system has the miss recognition of the user's speech.

We have developed an anthropomorphous agent system called "Visual Software Agent" or simply "VSA"(Dohi & Ishizuka 1993; 1996). The VSA has a realistic 3-D face as well as a speech dialog function. Its face has a texture image obtained from a real human and is always naturally rocking; these features give friendliness and sometimes fascination to users. Recent computer graphics technology makes possible to generate a vivid realistic facial image and change its facial expression in real-time. It is no doubtful that the realistic face close to a real human face is better than a simple 2-D animation.

Here we combine our visual anthropomorphous agent VSA with the World-Wide Web (WWW) (Berners-Lee *et al.* 1994). The advantages of this combination are;

1. the Hyper-Text Markup Language (HTML) description identical to the WWW pages can be used for providing information through the VSA, and
2. the VSA can use vast information stored in the WWW and information services built in the WWW. As a result, hyper-media data in a standardized format can be used for the VSA as well as for the WWW.

Moreover, the VSA can play a navigator or guide with a friendly face and a speech dialog function for the WWW in the Internet.

Figure 1: The Visual Software Agent (VSA)

Visual Software Agent (VSA)

Figure 1 shows an operation example of the VSA with a woman's face. The VSA is emerged on the right display. A user puts a headset on and talks with her. She (the VSA) looks at the user (turns her head), winks, and speaks.

The VSA's face is formed with approximately 500 surface patches (triangle polygons), and a real human facial picture is transcribed onto it. The real-time deformation of the triangle polygons generates the change of facial expressions. The VSA winks his/her eyes, opens his/her mouth like speaking synchronized with a speech synthesizer. Since the real face alive is always rocking and never freeze, our VSA rocks his/her face naturally all the time. The rocking face gives intimate feeling that can not be obtained from static images.

For speech communication, two speech synthesizers and a speech recognizer are attached to the workstation through RS–232–C. Speech synthesizers are the "DECtalk" (Digital Equipment Corp.) for English and the "Shaberin-bo" (NTT data) for Japanese.

The speech recognizer is the "Phonetic Engine 200" (Speech Systems Inc.) with Japanese speech-recognition function as an addition. It is a speaker-independent system, and can recognize continuous speech.

VSA–WWW/Mosaic Connection

The combination of the VSA with the WWW/Mosaic has two meanings.

1. VSA accesses WWW servers as vast information databases, and then uses those data
   The WWW is a world-wide distributed information database. It has large amount of information which covers wide area, including multimedia data like images and sounds. Its information is always updated by the united efforts of a great number of people. Thanks to the VSA connection with the Mosaic, hyper-media data in a standardized format can be also used for the VSA. The VSA can access the WWW server directly, or through the Mosaic as a hyper–text viewer.
2. VSA offers another communication channel for the Mosaic
   On the Mosaic display, clicking a mouse on the highlighted string (called "anchor") invokes the retrieval of a linked object. Regardless of the type of WWW servers, it has unified operation interface. However all of users can not manipulate a mouse satisfactory. In other words, it may be unusable for people unfamiliar with a computer and physically handicapped persons. Also, there are situations that the mouse interface is not suitable, for example, in a system using a wall-type room-wide display. With the VSA connected with the Mosaic, a user can walk around the Internet world with speech dialog in addition to a current mouse interface on the Mosaic. Both the speech dialog operation through the VSA and the mouse operation have the same priority. All mouse operations are available any time independent of the use of speech dialog operation.

VSA–Mosaic Communication

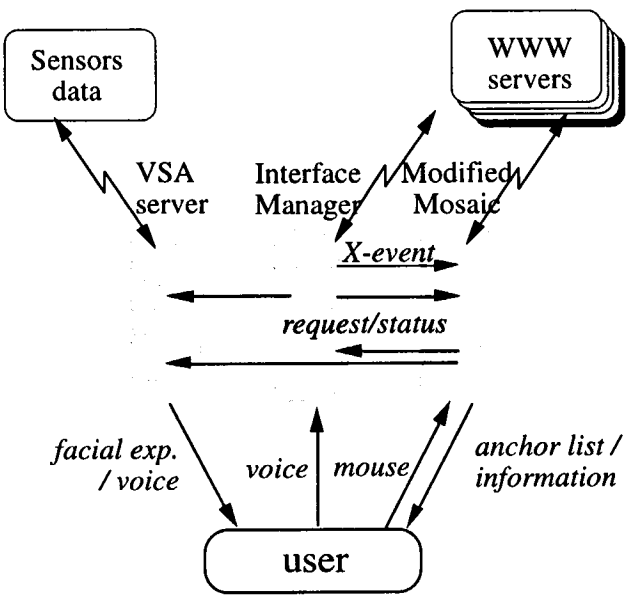

Figure 2: The design of VSA–Mosaic communication

Figure 2 illustrates the design of a VSA–Mosaic communication.

The following three independent processes communicate each other and collaborate. These processes can work on different workstations connected through TCP/IP.

- Interface Manager (speech recognition and a local database)
- VSA Server (facial image synthesis and speech synthesis)
- Modified Mosaic

Interface Manager The Interface Manager includes a speech recognition module, a keyword table, and a local database.

The speech recognition module recognizes speaker–independent and continuous speech. Because of the restriction of our speech recognizer, we will describe sentence-structure rules and word-category definition usage for speech recognition. The speech recognizer returns both the sentence recognized and its score for decoding of the complete utterance. The score value is between 0 to 999 (best). If the score value is under a threshold, the VSA system discards the returned sentence and prompts the user to speak the sentence again.

A keyword string in the speech sentence is usually passed to the modified Mosaic process as an anchor string for selection. The user can change the action associated with the keyword string by the keyword table. The keyword table assigns a keyword to a URL (Uniform Resource Locator) or an anchor string. The basic format of the keyword table entry is as follows.

$$[\ keyword\ \ [\ URL\ |\ \langle\ anchor\_string\ \rangle\ ]]$$

The first argument is the keyword appeared in the speech sentence, and the second argument is the URL or the anchor string passed to the modified Mosaic process.

In the case of an anchor string, it must be enclosed in "⟨" and "⟩". A space character between "⟨" and "⟩" also has a meaning.

A maximum-length matching search is used for keyword finding. Thus, the string "*foo bar*" has priority over the word "*foo*". When the second argument is omitted, the keyword string is used as the anchor string.

A local database keeps information converted into a suitable format for the VSA. It has also accumulates WWW server access histories.

The interface manager communicates with a user using speech dialog, and delivers a user's voice request to the Mosaic. It also sends speech texts for voice reply to the VSA server.

VSA Server The VSA server controls both realistic facial image synthesis and speech synthesis in order to realize simple lip sync.

When the VSA server receives a speech text from another process, it generates vowel sequences from the text. The VSA has six mouth shape patterns; five Japanese vowels ('a', 'i', 'u', 'e', and 'o') and close. Then it controls a facial image synthesis module along with vowel sequences, adding some supplemental motion parameters (wink eyes, rocking face, turning head, etc).

We have developed so far three facial image synthesis modules; using

1. original small-scale parallel transputers,
2. motion JPEG method on a JPEG decompression board, and
3. a fast graphics board with hardware texture mapping facility.

All modules have the same interface; therefore we can choose one as we need.

Modified Mosaic To make use of the Mosaic as a hyper-text viewer for the VSA, we have modified the source code of the NCSA Mosaic–2.4–L10N (Localization). It can handle HTML pages including Japanese characters well. An X–window event handler routine is added for the VSA–Mosaic communication. The Mosaic–2.5 or later, have the Common Client Interface (CCI) facility for an inter-process communication, but we don't use it because we have attentive control.

The Interface Manager process creates an X–window event structure when the speech recognizer detects one of keywords, and sends it to the Mosaic process. It informs the Mosaic process the occurrence of the request from a user. Then some requests/statuses are exchanged between the Interface Manager and the Mosaic through an X–Cut–Buffer. The modified Mosaic generates an anchor table on each page. If the keyword, received from the Interface Manager, is the anchor string, it searches the anchor table. And then if matched, it opens the linked page.

Speech Dialog Operation

The current VSA recognizes four types of speech commands for Mosaic operations. Below are these command samples, which depend on sentence structure rule definitions of the speech recognizer.

- Select–by–Keyword

  ex.) "Please show me *XXXX*."
  "*XXXX*, please."

  If *XXXX* (or associated anchor string listed in the keyword table) is matched an anchor string in a WWW page, it opens its linked page. This is equivalent to click with a mouse on the anchor string.

  If the second argument of the keyword table entry is the URL, it jumps to the URL page. This is equivalent to select the title and then jump to the page using on a hotlist (or bookmark) function.

- Select–by–Index–number

  ex.) "Please show me number *N*."
  "Number *N*, please."

Whenever new page is opened, all anchors on the page are listed with index numbers automatically. This speech command enables the selection of the *N*-th anchor on the list. This function is prepared for anchors without character strings like image anchors.

The restriction of the speech recognizer is another reason. We must provide sentence-structure rules and word-category definition usage to the speech recognizer. It is apparently impossible to register all anchor strings on WWW servers into the speech recognizer.

- Select–by–URL

  ex.) "Please contact (or connect) with *XXXX* server."

  If *XXXX* is listed in the keyword table, it jumps to an associated URL page directly.

- Misc. commands
  These are several reserved words such as;

  ex.) *"Home page"*
  return to the home page
  "forward"/"back"
  forward/back the page link
  "page up"/"page down"
  page up/down on the page.

  These are equivalent to click certain functional buttons or a scroll bar on the Mosaic display.

Additional Intelligent Functions

Sub Agent

If a boss has a habit to ask every morning, "how is the weather today?", a capable secretary checks a weather forecast on a newspaper beforehand and replies immediately, "a newspaper says it will be rain in this afternoon, so you should take your umbrella with you."

Current typical interface style on the WWW is "on demand." When you want to get any information, you clicks the mouse button and get information. However, it doesn't start displaying information, but only loading information from the WWW server. Accessing the WWW server may sometimes take long time because of network traffic jam.

In the VSA system, a sub agent autonomously gathers information into a local database. It accesses periodically specific WWW servers, then picks necessary information out and converts it into a suitable format for the VSA. For examples, they may be weather forecast, stock prices, and relevant news topics which include specific keywords, etc. Since a page layout is usually fixed even if articles and values are frequently updated, it is not difficult to pick necessary information out from HTML texts.

In our system, some sub agents work every day. For example, one agent accesses through one company's WWW server three times per day and thus it keeps the latest weather forecast information from the Meteorological Agency Japan. The HTML doesn't provide a rigid page layout; how to offer information to users. Therefore retrieved information is converted from the HTML format into suitable format for the VSA; Thus the VSA can reply information in voice.

Find a Suitable Page

As the WWW grows explosively, it becomes difficult for a user to find a suitable page location. The WWW includes much junk data.

On the Mosaic, a user must select the title by a mouse to jump to an interesting page. Its interface style is a "direct manipulation." Its entry remains unchanged unless a user modifies explicitly a hotlist table.

Our basic idea is that the system records access histories of users and shares / exchanges them among neighbor users or users in the same group. A priority value is adding to the pair of a title/keyword and its URL. Each entry on the history database is a triplet,

$\langle keyword, URL, priority \rangle$.

Large priority value has a priority. We assume a user will access interesting pages frequently and stay on long time. If the user accesses the page, it increases its priority value. All priority values are periodically decreased. Thus the history database changes gradually. The user may merge neighbor's histories into his/her own database. Its change also depends on the choice of "neighbors".

Our strategy on selecting pages is as follows.

1. When a user asks about a keyword, the history database is checked up. If the keyword entry is included in the database, candidates are presented to the user. Then if the user chooses one of those, its priority value is increased.
2. If 1.) is missing, the system sends queries to some neighbor agents. (Or the system may peep into neighbor's history database if it is exported)

   It is easy to exchange and share a URL link pointer. Since it is the location of information, instead of its contents, it requires only small memories. It depends on each agent how to handle the contents.
3. If the keyword is found on neither own nor neighbors' history databases, it is passed to a certain internet search engine.

Communication Examples

Figure 3 (appeared in last page) shows an example display of the VSA–Mosaic interface. There are three windows on a screen. The right window shows a modified Mosaic page. Its appearance is the same as the original Mosaic system. The upper left window is the VSA with a rocking realistic face. In the lower left window, it displays the anchor list whenever new page is opened.

Example 1 below shows a simple dialog example. A user speaks to the VSA, and the VSA replies in voice as well as changing the Mosaic page, if necessary.

User: Please contact with "*the University of Tokyo*" server.

VSA: Yes. Just a moment please.
(connect with the server, then present an anchor list)
O.K. There are 5 anchors.

User: Number 0, please.
(0: Japanese version)

VSA: (open "*Japanese*" page, and present new anchor list)
O.K. There are 5 anchors.

User: Please show me c????s ...
(fail speech recognition)

VSA: Pardon?

User: Please show me "*campus map*"

VSA: (open "*campus map*" page, and present new anchor list)
O.K. There are 3 anchors.
...

Example 1. A simple dialog example.

Evaluation

Currently our target users are not computer experts. The realistic human face in the VSA system attracts strongly the user's attention to the display. "Rocking" is also very important factor. However it may be an annoying interface for computer experts. They are accustomed to computer operations (key typing, mouse operation, etc.) well. The speech input can not be always recognized correctly because the performance of any speech dialog systems is insufficient yet for free talk.

We don't intend to replace all mouse operations with the speech dialog operations. The speech dialog operation is a complementary method. It is important to choice the most suitable operation method in accordance with various situations.

If all anchor strings have unique names each other, the speech dialog operation has the advantage, because the user need not to move a mouse cursor onto the anchor string. For example, when the user checks all pages by the mouse operation one by one, the mouse cursor goes back and forth between the "BACK" button and anchor strings.

And the speech dialog operation doesn't need any device nor any exercise. The mouse operation may be unusable for people unfamiliar with a computer and physically handicapped persons. Also, there are situations that the mouse operation is not suitable, for example, in a system using a wall-type room-wide display.

On the contrary, the speech dialog operation can't necessarily point location out directly because of the characteristics of speech. In the case there are more than two identical anchor strings on one page, the mouse operation can distinguish those easily, but the speech dialog operation can't. Two anchors with the same keywords don't necessarily have same links.

The use of non-string anchors like image anchors is one of the special features of the HTML. However the speech dialog operation can't handle these anchors well, because the image anchor has no anchor string. In the VSA system, the user can use added index numbers instead of anchor strings.

Conclusion

In this paper, we have described a prototype of the Visual Software Agent (VSA) system connected with the WWW/Mosaic. The VSA offers a new multimodal interface with a realistic moving facial image and speech communication to access the WWW as a vast information database. The VSA also has some sub agents, some of which access periodically specific WWW servers, and converts gathered information into a suitable format for the VSA. In our current system, one sub agent is working everyday to keep the latest weather forecast.

To facilitate finding suitable pages on the WWW, the VSA maintains an own access history database, and exchanges a triplet $\langle keyword, URL, priority \rangle$ among neighbor agents. It is easy to share triplets because it is a pointer to information location, not contents itself. Priority values are modified by accessing pages and by importing neighbor history. Candidate URLs relevant to a keyword are changed automatically. If candidates don't fit user's request, the keyword is passed to a certain internet search engine.

The VSA offers a user another communication channel, an interface style close to daily face–to–face communication with speech dialog. It delivers a user's speech request to the WWW/Mosaic. A user can walk around the Internet by speech dialog with the visual anthropomorphous agent, VSA. It gives users friendliness feeling. It is not only useful for little kids, persons unfamiliar with a computer and physically handicapped persons, but also may broaden the usage style of the WWW information and functions. The interaction between the agent and a user becomes more smooth and natural and the agent can navigate the user in the vast Internet information space.

References

Berners-Lee, T.; Cailliau, R.; Luotonen, A.; Nielsen, H.; and Secret, A. 1994. The World–Wide Web. *Commun. of ACM* 37(8):76–82.

Dohi, H., and Ishizuka, M. 1993. Realtime Synthesis of a Realistic Anthropomorphous Agent toward Advanced Human-Computer Interaction. In Salvendy, G., and Smith, M., eds., *Human-Computer Interaction: Software and Hardware Interfaces*, 152–157. Elsevier.

Dohi, H., and Ishizuka, M. 1996. A Visual Software Agent connected with the WWW/Mosaic. *Int'l Symp. Multimedia Systems (MMJ '96)* 392–397.

Etzioni, O., and Weld, D. 1994. A Softbot–Based Interface to the Internet. *Commun. of ACM* 37(7):72–76.

Etzioni, O., and Weld, D. 1995. Intelligent Agents on the Internet: Fact, Fiction, and Forecast. *IEEE Expert* 10(4):44–49.

Laurel, B. 1990. Interface Agents: Metaphors with Character. In Laurel, B., ed., *The Art of Human–Computer Interface Design*, 355–365. Addison–Wesley.

Maes, P. 1994. Agents that Reduce Work and Information Overload. *Commun. ACM* 37(7):31–40.

Nagao, K., and Takeuchi, A. 1994. Speech Dialogue with Facial Displays: Multimodal Human–Computer Conversation. *32nd Annual Meeting of the Association for Computational Linguistics (ACL–94)* 102–109.

Takeuchi, A., and Nagao, K. 1993. Communicative Facial Displays as a New Conversational Modality. *ACM/IFIP INTERCHI'93 (INTERACT'93 and CHI'93)* 187–193.

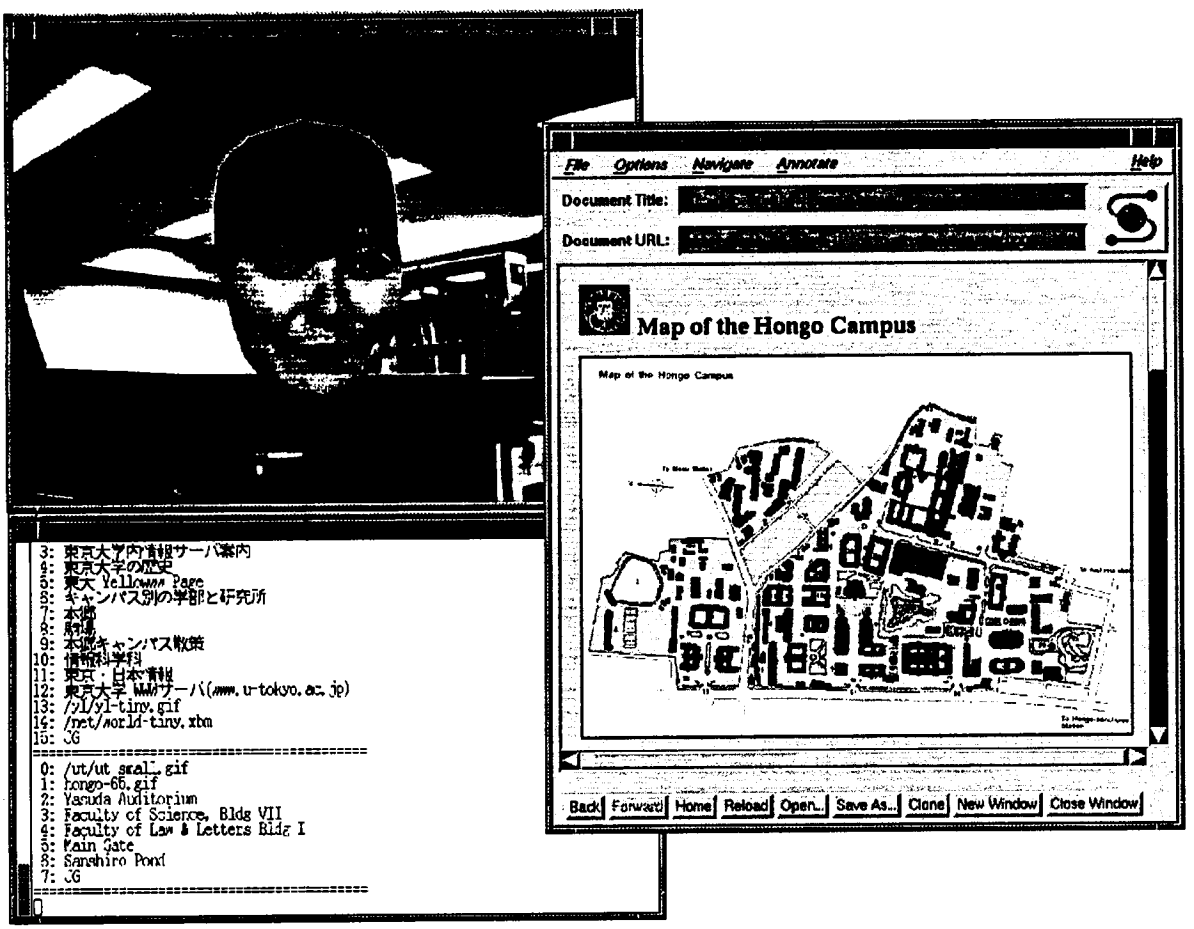

Figure 3: VSA–Mosaic interface