Info-concierge: Proactive Multi-modal Interaction through Mind Probing

Takatsugu Hirayama,* Yasuyuki Sumi,† Tatsuya Kawahara,‡ and Takashi Matsuyama§
* Graduate School of Information Science, Nagoya University, Aichi 464-8603, Japan
E-mail: hirayama@cmc.ss.is.nagoya-u.ac.jp
† Future University Hakodate, Hokkaido 041-8655, Japan
E-mail: sumi@acm.org
‡ Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606-8501, Japan
E-mail: kawahara@i.kyoto-u.ac.jp
§ Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
E-mail: tm@i.kyoto-u.ac.jp

***Abstract*—To close the wide gap between the exploding information world and individual human knowledge, we have investigated a multi-modal interaction strategy called *Mind Probing* in which a system proactively acts on the user and then estimates his/her internal state by analyzing his/her reaction. We have also developed a system, *Info-concierge*, that can probe the latent interest of the user, make him/her aware of it, and proactively provide sensible information. The system uses the core techniques of *Mind Probing* through speech and gaze recognition. In this paper, we present the interaction flow, the core techniques, and a field trial of *Info-concierge*.**

I. Introduction

People are inundated with enormous volumes of information and face difficulty in finding the information they desire. To get this information, they need to have a clear target and to master a keyword/command-based interface. Otherwise, they might become frustrated. The problem has been observed among not only particular generations but also whole generations. That is to say, the digital divide has become widespread. Therefore, interactive support by a well-informed agent such as a concierge of a luxury hotel might effectively alleviate their concerns (see Fig.1). Our work aims at creating a concierge system that can probe the latent interest of a user, make him/her aware of it, and proactively provide sensible information. We call this system *Info-concierge*. Based on interaction with the system, the user can explore something that is on his/her mind and arrive at a target that satisfies his/her interest and information demands.

Conventional interactive systems (e.g., MIT VOYAGER [1], ALICE [2], and existing web search engines) use a reactive interaction strategy. These reactive systems can respond only to specific commands from the user. This strategy is not effective when the user is not aware of his/her own internal state (such as interest or intention). To go beyond reactive interaction, some researchers have proposed statistical analysis methods to estimate the user's internal state by passively sensing subconscious or non-verbal behaviors [3]. However, it is difficult even for a human to estimate internal states using only passive sensing.

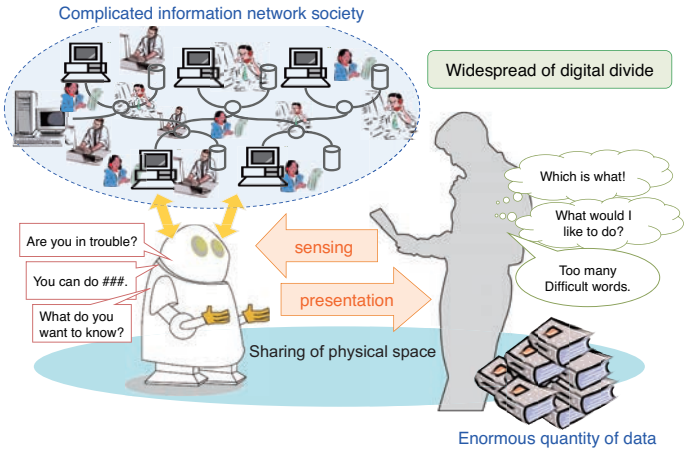

Fig. 1. Human computer interaction in an era of information explosion. Info-concierge is a bridge between the information world and the physical world to eliminate the digital divide.

How do people infer the internal state of others? Our research is inspired by basic scenes of interaction that occur in everyday life. To understand the conversational partner's internal state, people proactively act on the partner via verbal and non-verbal behaviors, such as by bringing up general topics or establishing eye contact, and thereafter sense the partner's reaction; proactive approaches encourage the partner to reveal his/her internal state through reply timing, gaze behavior, and so on[4]. Using the analogy of proactive interaction, the system should elicit the user's reaction on its own initiative through multi-modal interaction, and estimate his/her internal state by analyzing his/her reaction. We call this internal state estimation strategy based on the proactive interaction *Mind Probing*. In this study, we incorporate some sorts of methods based on Mind Probing into Info-concierge. All of them are the result of the "New IT Infrastructures for Information Explosion Era (Info-plosion)" research project funded by the MEXT Grant-in-Aid for Scientific Research in Priority Areas.

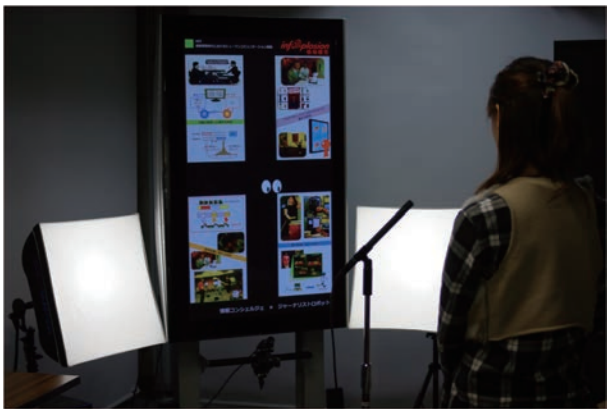

Fig. 2. Info-concierge with a large screen display. The system presents an anthropomorphic agent (in the shape of eyes in the center area of the screen), four pieces of content in the corners, and audio information. The system estimates the user's interest and intent based on the gaze and speech reactions to the dynamic presentation.

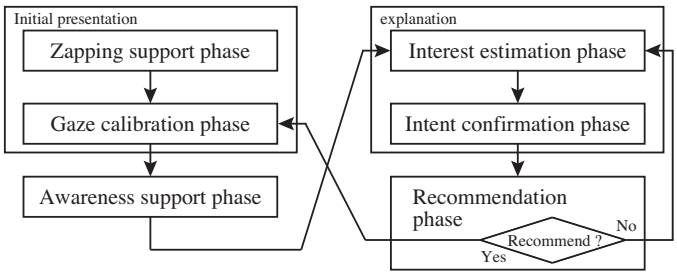

Fig. 3. User-system interaction flow. The flow consists of six phases. Info-concierge interacts with user based on Mind Probing in each phase.

II. Overview of Info-concierge

A. Setup

Info-concierge is a bridge between the information world and the physical world, which is necessary in various scenes in daily life, e.g., as a guide for travelers, a supporter of drivers, a facilitator in family meetings, and so on.

This study assumes a scenario setup as shown in Fig.2. The system presents some visual content on a large screen display and supplementary audio information via a speaker; the user compares the pieces of content and selects a preferred one (under his/her knowledge constrains). We humans often face a situation in which we are presented with many news articles and wish to make a choice (e.g., on yahoo.com, amazon.com, and so on); however, most of us cannot make this decision easily. We believe that the interactive support provided by the system will be helpful to a hesitant user. If the system provides some detailed information or recommendation adapted to the user's interests, it can reduce his/her ambivalence. To achieve effective support and natural interaction with a novice user, the system should sense his/her interest using non-intrusive devices [5], namely a camera and a microphone; and employ an anthropomorphic agent who provides natural, appropriate information on the screen [6], [7]. In this work, we attempt to estimate which content the user is interested in by eliciting user's gaze and speech behaviors.

B. Content

Info-concierge holds many pieces of content within a certain field. Each piece of content consists of several pages worth of articles and textual explanations. The system displays up to four pieces of content on the screen at the same time (the screen is divided into four equivalent areas) and outputs the textual explanation with a text-to-speech synthesis program. The agent is displayed in the center area of the screen. All content is classified into several categories based on their similarities. The calculation of similarity is based on the matching of keywords included in each piece of content.

The Info-concierge system was installed at the entrance of a poster presentation venue at a research conference; this provides an example of implementation. Fifty research topics on human communication were presented in the form of posters at the venue. These topics were classified into eight categories based on their similarities: interaction dynamics, proactive interaction, speech recognition, motion and behavior recognition, natural gesture synthesis, human-human communication support, life activity support, and multi-party communication analysis. Info-concierge estimated which research topic the user was interested in via a multi-modal interaction, and recommended related topics or guided the user to a specific poster presentation site.

C. Interaction flow

The interaction flow consists of six phases: zapping support, gaze calibration, awareness support, interest estimation, intent confirmation, and recommendation phase, as shown in Fig.3. The details of the interaction techniques are shown in Section III.

1) **Zapping support phase (initial presentation phase):** When people make a choice from various pieces of content, most of them will first glance over the entire set of available choices. They cannot easily request anything from information service systems without knowing what they are provided with. We therefore introduce a proactive interaction that supports zapping (a process in which the user quickly changes his/her attention) of the user to Info-concierge. In the first phase, wherein the system proactively enumerates the eight research categories, the user selects an attractive category and indicates it using speech. The system recognizes the utterance's intention in view of the user's response timing during the enumeration [8].

2) **Gaze calibration phase (initial presentation phase):** After the zapping support, the system presents four representative research topics from the chosen category. In the second phase, the system shows the first article for each topic, and at the same time, calibrates spatial parameters to compute the user's gaze position on the screen. Gaze tracking is an important process to monitor if one is to understand a user's internal state. To improve the tracking accuracy, the system

displays the four articles in one area after another and draws the user's gaze reactions to the display events [9]; it then derives a mapping between the display areas and the spatial directions of the reactions.

3) **Awareness support phase:** At the beginning of the interaction, too much proactivity from the system may wilt the user. The system needs to attract the user and elicit his/her interest subconsciously for a while. We apply synchronous imitation of a user's gaze behavior to make latent interest explicit in the third phase. This imitation makes the user aware of his/her own interest, that is, it supports the self-awareness of interest [10]. The system can roughly estimate which research topic the user is interested in because the imitation elicits the user's gaze behavior, which reflects his/her interest.

4) **Interest estimation phase (explanation phase):** The system then talks about the topic of the interest in detail. During this time, the user does not always pay attention to the explanation of the topic. When the user is not engaged in the conversation on the topic, the system should change the topic. We focus on joint attention, i.e., keeping the user's visual attention on the content corresponding to the agent's speech explanation, as a behavior displaying engagement. The system estimates engagement using the correlation between the system's utterances and the user's gaze behaviors [11]. Next, the system determines whether or not the user is interested in the content of the agent's speech explanation. When the user is not engaged with the explanation, the system updates the visual content of the explanation and observes the user's gaze reaction; it then estimates interest using the timing structures between content-display updates and gaze reactions [12]. The system changes the explanation topic if the user does not quickly react. Otherwise, the system continues to talk about the topic.

5) **Intent confirmation phase (explanation phase):** It is not easy to estimate interest using only the user's gaze reactions. To elicit an overt response reflecting the user's internal state, we focus on the agent's gaze behavior at interactionally significant places [13]. Specifically, the system turns the agent's gaze toward the user at the point at which it finishes a sentential unit when questioning or recommending. In this manner, the system confirms the elicited intent based on the response.

6) **Recommendation phase:** After a series of proactive interactions based on Mind Probing, the system determines whether or not to recommend four research topics related to the user's interest. In the previous phase, the system asked the user whether s/he would like to proceed to detailed information regarding the topics after the system had finished the explanation of a research topic. If the user does not quickly respond to the question, the system proactively acts on the user by casually asking again or giving additional information on the topic. If the user immediately accepts the offer, the system presents the related topics by transiting to the gaze calibration phase, otherwise it explains the remaining article topics on the screen by transiting to the interest estimation phase.

III. Mind Probing-based interaction in Info-concierge

A. Intent recognition using timing of user response during enumeration of content [8] (zapping support)

In human-agent interaction systems, users prefer to be able to speak at anytime and use natural expressions. However, users cannot easily request anything if they do not know the information content on the system. In this case, a proactive presentation of the samples, i.e., a subspace of information on the system, to the user is effective. We have designed an interaction in which a user makes a choice from various pieces of content while the system enumerates them one by one.

The system's ability to read out each piece of content from a list is important for two reasons. First, the user can indicate a choice via timing information, which can be detected robustly. The results for barge-in timing are more reliable than automatic speech recognition (ASR) results in many cases. Second, this interaction often appears when a system displays retrieval results in the information retrieval task; this is a promising task in the conversational dialogue systems developed at several companies such as Microsoft [14] and Google[1].

For example, the system and the user might interact as follows:

> **System:** There are eight categories that I would suggest. "Research on interaction dynamics", "Research on proactive interaction", ...
> **User:** That one.
> **System:** OK, you mean "Research on proactive interaction". First, I'll present four research topics in that category.

In this example, the user barges into the utterance of the system while it reads out "Research on proactive interaction". The system identifies the user's referent, that is, what s/he indicates by "That one". The system can recognize that the user selected "Research on proactive interaction" by focusing on the barge-in timing of the user utterance.

Matsuyama et al. [8] investigated the timing distribution of user utterances containing referential expressions. They defined barge-in timing as the time difference between when the system utterance starts and when the user utterance starts. As a consequence, the average barge-in timing, $\bar{T}_{bargein}$, was 1.2 seconds when the system enumerated content titles with an average utterance time, $\bar{T}_{enum}$, of 0.73 seconds and a pause time, T_{pause}, between the enumerated content of 1.0 seconds. On another condition, the relationship was $\bar{T}_{bargein} = 2.2$ seconds for $\bar{T}_{enum} = 5.27$ seconds and $T_{pause} = 2.0$ seconds. That is, the user's utterance timing has a high correlation with the temporal parameters of the system's utterance. They postulated that the users needed to listen to at least some portion of the system's utterance regarding the target content before s/he would decide to select it.

[1] http://www.google.com/goog411/

The users utter not only referential expressions, but also content expressions containing the content title such as "Research on proactive interaction". They often use the referential expressions when the enumerated content title is long (in time) or contains unknown words. On the other hand, they tend to use content expressions when the enumerated content title is short or when they are barging in to indicate previous system's utterances. We therefore integrate the two different information sources for barge-in timing and symbolic ASR results to recognize the user's intent.

In practice, Info-concierge visually and aurally enumerates the eight research categories at T_{pause}-second intervals. We estimate $\hat{T}_{bargein}$ from the value of $\bar{T}_{enum}$ of their titles based on the above correlation, and allow the user to indicate a category i using the referential expression from $t_{i,begin}+\hat{T}_{bargein}-\sigma$ to $t_{i+1,begin}+\hat{T}_{bargein}-\sigma$. Here, $t_{i,begin}$ denotes the beginning time, the time at which the system begins to enumerate category i; σ is empirically set based on individual variation.

B. Gaze Probing: Event-based estimation of objects being focused on [9] (gaze calibration)

Knowing which content the user turns his/her visual attention (eye-gaze) toward on the screen is crucial to understanding his/her interest. Most existing eye-gaze tracking methods need to learn a mapping between artificial visual targets and gaze directions (including calculation error), what is called gaze calibration. This is a troublesome and unnatural interaction with the user for Info-concierge to engage in. To realize subconscious and stable calibration, we apply *Gaze Probing*, which is based both on a dynamic presentation of content, and on a timing measurement of the user's gaze reactions, to gaze calibration at the beginning part of interaction.

Gaze Probing was proposed by Yonetani et al. It is an event-based method for estimating the target object of the user gaze using "designed dynamic content" [9]. They used the synchronization of motion between objects and the user eyes as a cue. The majority of existing eye-gaze tracking methods perform a direct comparison between the positions of the objects on a screen and the user's eye-gaze directions. Estimating an eye-gaze direction requires a tradeoff between accuracy and the user's freedom of movement. Active sensing techniques, e.g., the pupil center corneal reflection (PCCR) technique using an infra-red camera and a light source, can obtain more accurate estimation, but apply some constrains to the user's head orientation and position. On the other hand, reactive sensing techniques that involve the use of a visible camera allow freer user movement in exchange for a larger margin of error. The latter is more appropriate in our assumption.

In the original formulation of Gaze Probing, an event is defined as a characteristic translation pattern of visual content. The event is embedded in the dynamics of content. The pattern basically consists of stopping within a certain time and scrolling in the horizontal direction with a constant speed. When content begins to scroll, the movement causes an expressive gaze reaction. Gaze Probing evaluates the synchronization between events and gaze reactions; the evaluation itself is based on the temporal distance between the starting point of each event and that of each eye movement.

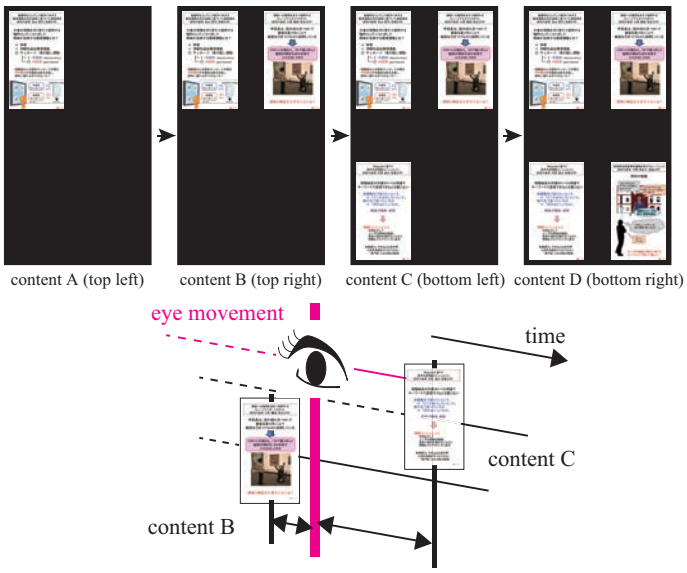

Fig. 4. Gaze Probing. The bottom figure shows the timing structure between the pop-ups of the articles and the gaze reaction, and that the user is focusing on content B displayed in the top right area of the screen.

Info-concierge presents four articles (the first from each of the four topics) on the screen, all of which belong to the research category selected by the user during the zapping support phase. The articles are displayed in one area after another, from the top left to the bottom right at an interval of $T_{interval}$ seconds. We regard the pop-up of the article as an event; we consider the user's gaze reaction to the event as the trigger for gaze calibration (see Fig.4). That is, the system detects the largest eye movement within $T_{gazeprobe}$ seconds after the event. The calibration is based on a mapping between the centroid of the content area and that of several gaze positions computed using a reactive sensing technique after the trigger. As the system displays four pieces of content at the same time, an affine transformation from four centroid sets is derived as a mapping function.

C. Gaze Mirroring: Imitation of gaze behavior for making user's latent interest explicit [10] (awareness support)

Proactive interaction is considered a core technique in Info-concierge. However, a very 'busy' approach may be annoying to the user at the beginning of the interaction (This is often experienced when one is abruptly approached by salesclerks right after entering a store). We believe that the system must 'keep its eye' on the user's behavior and attract him/her before asking questions. Also, it is more socially acceptable make the user's interest more explicit before explaining in detail or making recommendations.

Park et al. proposed a human-agent interaction called Gaze Mirroring that makes a user's latent interest explicit [10]. It is an imitation of the gaze behavior between a user and

Gaze tracking result (This image was captured by the camera located under the screen.)

Fig. 5. Gaze Mirroring. The anthropomorphic agent turns the gaze toward the user's gaze article (In this figure, the article is the top right one).

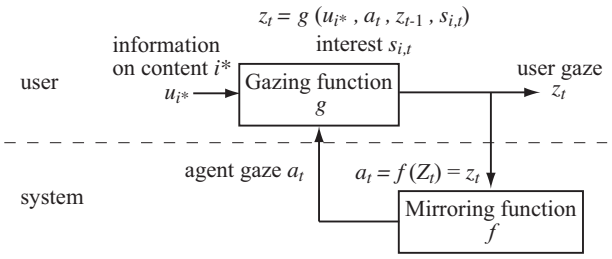

Fig. 6. User gaze behavior model based on Gaze Mirroring.

an anthropomorphic agent. In other words, Gaze Mirroring is a kind of biofeedback. The agent turns its gaze toward the object of the user gaze synchronously (see Fig.5). The imitation makes the user subconsciously aware of his/her own gaze behavior.

The user may establish joint attention with the agent through Gaze Mirroring. Joint attention is a process in which one shares an object of interest with others and understands others' mind [15]. The user may also be involved in joint attention with him/herself because the agent acts as an avatar, much like a mirror image, of the user. Therefore, the understanding of others' mind through joint attention can be transformed into self-understanding. In this way, it is possible to make the user aware of his/her own latent interest.

If the user interacts with the system as s/he would in human-human communication, joint attention will be able to influence the user's behavior. Park et al. demonstrated the effects of Gaze Mirroring experimentally: they elicited gaze behavior reflecting interest. The subjects gazed longer at an object of their interest. In cases in which the users were gazing at their object of interest, they would not feel stressed by the imitation. Thus, the user felt affinity with the agent and continued joint attention with the agent. On the other hand, in the case of no interest, the users would turn their gaze away to suggest their indifference. The interaction model is outlined in Fig.6. Gaze Mirroring is either executed for T_{mirror} seconds or is aborted when the accumulated duration of gaze at a piece of content is sufficiently larger than for the other content, e.g., when it exceeds 50% of the execution time. Info-concierge roughly estimates the user's interest based on the accumulated duration, and then begins talking about the topic of interest.

D. Dialogue engagement estimation using correlation between system utterances and user gaze behaviors [11] (interest estimation)

In face-to-face conversations, speakers sometimes glance at a listener and check whether the listener is engaged in the conversation. Listeners display their engagement through verbal/nonverbal behaviors such as back-channeling and eye contact. When the listener is not fully engaged in the conversation, the speaker changes the conversational content or strategy.

This engagement checking process is fundamental and indispensable not only in face-to-face conversation, but also in human-agent communication. If the communication channel between the user and the agent is not well set, information presented by the system will not be properly conveyed to the user. If the system can monitor the user's attention to the conversation and detect whether the user is engaged or not engaged in the conversation, then the system can adapt its behaviors and communication strategies to the user's state. For instance, if the user is not engaged in the conversation, the system may need to attract the user's attention by changing the conversation topic.

Nakano et al. [11] collected a conversation corpus and made a subjective evaluation of the degree of engagement in the user-agent communication to find the engagement and disengagement patterns of users' behaviors. They focused on the users' gaze behaviors and analyzed 3-grams of gaze direction transition using 4 labels: looking at the content of the agent's speech explanation, looking at the agent's head, looking at the agent's body, and looking at the other content. The first label indicates whether joint attention is established between the user and the agent.

As the results of the analysis (including a comparison with the subjective evaluation), 3-grams with a lower degree of engagement did not have the first label, i.e., looking at the content of the agent's speech explanation, whereas those with a higher degree did not have the second and fourth label, i.e., looking at the agent's head and at the other content. This suggests that establishing the joint attention is an indispensable way of estimating a user's conversational engagement. We applied this basic theory to Info-concierge, which estimates engagement based on the duration of joint attention for a certain time period, T_{engage}. For instance, when the duration exceeds 50% of T_{engage}, the system judges the engagement as high (see Fig.7 (a)).

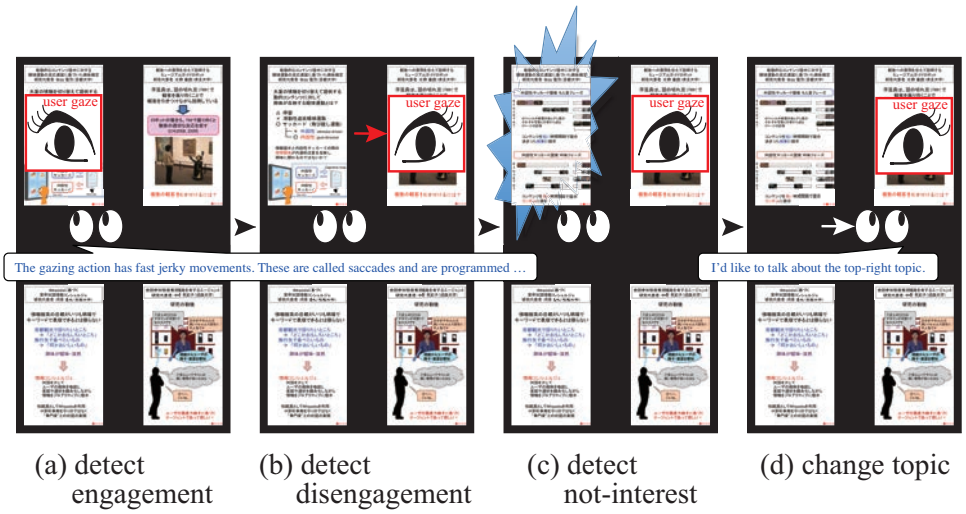

Fig. 7. Engagement and interest estimation based on Mind Probing.

E. Interest estimation using timing structures between proactive content-display updates and eye movements [12] (interest estimation)

After engagement estimation, the system needs to determine whether or not the user is interested in the content of the agent's speech explanation, so as to control the conversation topic. As a method of estimating a visual object of interest for a user, Hirayama et al. employed proactive content-display updates in which the system displayed the articles making up the content, one after another, on the screen [12].

They focused on the relationship between the dynamics of the content-display updates and the user's gaze reactions, and thereby defined two temporal features that relate to covert attention.

- *Reaction*: the response time to switch the gaze to the next update which occurs in a different part of user's visual field.
- *Resistance*: the duration the gaze fixes on the previously updated content, regardless of the next update. This has same value as *reaction*, but is defined in terms of the previously gazed content.

They made the following hypothesis. The *reaction* will be shorter or the *resistance* will be longer for the interesting content. Some experimental results support the hypothesis regarding *resistance*. They confirmed that under dynamic content presentation the temporal features can be more efficiently used to estimate the interest of the user than the conventional features such as gaze duration and frequency.

Based on their results, the user is not easily aware of the other content updates around the gaze content, if the content attracts strong interest. Even if the surrounding content update seems to cause exogenous eye movement, it will have almost no influence on the user's behavior. Info-concierge therefore switches the article of the agent's speech explanation when the user is not engaged in the explanation, i.e., when the user turns his/her attention to the surroundings of the article (see Fig.7 (b) and (c)). The system then switches the explanation topic to the user's gaze article if the user does not return his/her gaze to the article being explained within a certain number of seconds, $T_{interest}$, i.e., when the user is not interested in the article (see Fig.7 (c) and (d)). On the other hand, when the user is engaged in the explanation or immediately returns, the system continues to talk about the topic while estimating engagement and updating visual content frequently.

F. Intent elicitation by turning agent gaze at transition relevance places [13] (intent confirmation)

Mind Probing cannot completely estimate a user's interest because of the complexities of human mind, although it has more accurate performance than passive sensing-based methods. The concierge system needs to elicit user's real intent from the clue given by its rough estimation result. That is, the system needs to confirm which content the user would like to select; this should be done through proactive interaction. We focus again on proactive gaze behavior as a way to elicit a user's reaction, which reflects intent.

Kuno et al. [13] investigated how human guides coordinate their behavior with their talk when explaining exhibits to visitors with the aim of designing a museum guide robot. In particular, they analyzed at what points during the talk the guides turned their heads toward the visitors. Consequently, they noted transition relevance places (TRPs) which are the points at which a speaker is likely to hand over the turn to a listener, such as upon finishing a sentential unit. That is, TRPs are the most frequent point at which the guides turn their gaze toward the visitors.

The guide may be able to check the visitor's understanding or non-understanding by turning his/her head, as well as confirm whether the visitor is listening. In addition to TRPs, the guides turn their head toward the visitors when saying key terms. The head movement again allows the guide to check the visitor's visible displays of understanding.

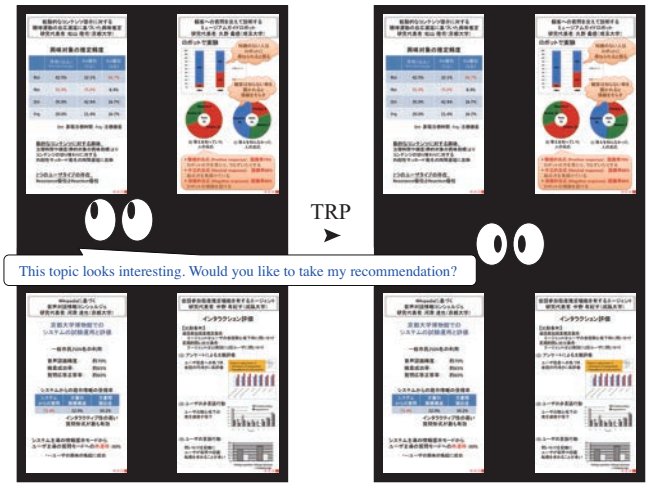

Fig. 8. Intent confirmation based on Mind Probing. The agent turns its gaze toward the user at TRPs (Transition Relevance Places).

Based on the results of their experiment using human guides, they developed a robot that moves its head while explaining two posters. Using the above interactionally significant points for head movements, they examined how visitors responded to the robot's head movements. They analyzed the experimental subjects' behaviors that started within one second of the robot finishing the turn of its head; they found two typical movements: nodding and mutual gaze. When the robot turned toward the subjects, they often made vertical head movements. The subject nodding may display an attempt to show understanding of the explanation. Furthermore, when nodding occurs at TRPs, it may function as a "continuer", or request to keep the explanation going. Second, subjects often gazed toward the robot from the poster almost as soon as the robot turned toward the subject (from the poster) during the explanation. This kind of behavior is regarded as mutual gaze as it seems to reveal the subjects' attempts to engage with the poster in concert with the robot. These behaviors increased when the robot turned its head toward the subjects at interactionally significant places, i.e., TRPs.

This result suggests that Info-concierge should elicit the user's overt response (reflecting an internal state) by this kind of proactive gaze behavior. The system therefore turns the agent's gaze toward the user at TRPs when questioning, e.g., "This topic looks interesting. Would you like to take my recommendation?" in this phase (see Fig.8).

G. Switching of recommendation strategy according to user's reaction [16] (recommendation)

During a period of proactive interaction flow based on Mind Probing, the system determines whether or not to recommend four research topics related to the most interesting (as determined by the system) topic on the screen. The four topics are extracted in accordance with a map of interest, which has distances based on the reciprocal of the similarities between the topics. The distances are weighted by degrees of non-interest for topics on the screen. The degrees are derived from

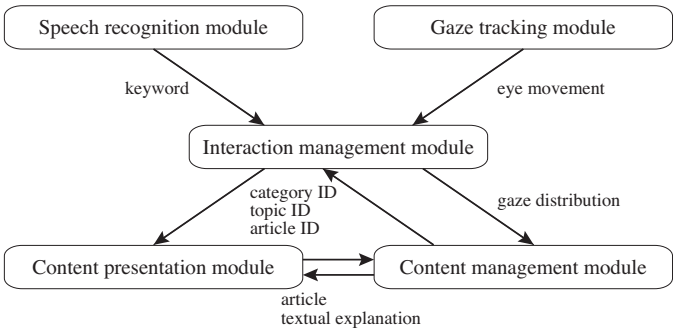

Fig. 9. Module organization.

a relative frequency distribution for the user not turning his/her gaze toward each piece of content in the awareness support phase and the interest estimation phase.

In the previous phase, the system asked the user about the recommendation after the system had finished its explanation of all of the articles of a research topic. If the user did not respond to the offer within $T_{recommend}$ seconds, the system proactively acts on the user either by asking a question such as "How do you like this?" to prompt his/her response, or by giving additional information on the topic. If the user immediately accepted the offer, the system presents the related topics by transiting to the gaze calibration phase, otherwise it explains the remaining articles (the other content) on the screen by transiting to the interest estimation phase.

IV. Implementation of Info-concierge

A. Module organization

Info-concierge is composed of five modules: speech recognition, gaze tracking, content presentation, content management, and interaction management module.

The interaction management module exchanges information with other modules (see Fig.9), and starts to communicate with the user when s/he appears in front of the Info-concierge; in concrete terms, when the gaze tracking module detects the face of a user. They interact at a distance of approximately 1.5 m. The sensing modules measure the user's utterances and eye movements using non-intrusive devices (a microphone[2] located in front of the screen and a camera[3] located under the screen) to achieve natural human-computer interaction. The content presentation module is composed of a large display[4] and stereo speakers[5]. The following sections explain each module in detail.

B. Speech recognition module

The Info-concierge recognizes user utterances using *Julius* [17], which is an open-source, high-performance speech recog-

[2]Sony electret condenser microphone ECM-23F5 and Edirol USB audio capture UA-1000, 44.1 kHz

[3]Point Grey Research IEEE1394b camera Grasshopper, UXGA, 30 fps, 8-bit gray

[4]Panasonic plasma TV TH-50PZ800, 50-inch, Full HD

[5]BOSE Computer MusicMonitor

nition software package for both academic research and industrial applications. It incorporates influential, state-of-the-art speech recognition techniques. For the input of a live audio stream, Julius first performs auto-splitting the input based on long pauses detected using energy and zero-cross thresholds. Julius then performs a large vocabulary continuous speech recognition task (LVCSR), which is processed effectively in real-time on low-spec PCs. It is also quite versatile and scalable.

We can easily build a speech recognition system by combining a language model and an acoustic model for the task, from a simple word recognition to a LVCSR task with tens of thousands of words. Julius supports various types of language model such as N-gram model, rule-based grammars, and a simple word list for isolated word recognition. Acoustic models should be of the Hidden Markov Model (HMM) type defined for sub-word units.

Applications can interact with Julius in two ways: socket-based server-client messaging and function-based library embedding. We apply the former to Info-concierge. In either case, the recognition results will be fed into the application as soon as the recognition process ends for an input. The Info-concierge can get the live status of, create statistics from, and control the Julius engine.

C. Gaze tracking module

The Info-concierge tracks the user gaze in four stages: face detection, estimation of face orientation, iris detection, and estimation of gaze direction.

First, the user's face is detected using the Intel OpenCV library, using Haar-like filters. Facial features (45 points) are extracted using the Active Appearance Model (AAM) algorithm [18]. The AAM is a statistical subspace model of shape and appearance. The system has an AAM trained using 150 face images of ten subjects. The face images of each subject were captured under 15 rotations of the head.

Next, a 3D face shape model is fitted onto the AAM by the bundle adjustment [19]; in fact, the translation and rotation parameters are optimized using the steepest descent method. The 3D model consists of 45 feature points, eyeball centers, and iris radius, which were measured using stereo cameras; their values are the average values for the ten subjects. The 3D position of the user's face is estimated as a result of the fitting. The irises are extracted by matching iris templates generated from the iris radius. Then, their 3D positions are estimated based on the eyeball centers and the iris radius.

After a straight line running through both the eyeball center and the iris center has been computed, its intersection with the display plane indicates the gaze position. The accuracy of gaze estimation is about 5 degrees (= about 10 cm on the screen) in real time (= about 30 fps[6]). The gaze position is corrected using the mapping function learned during the gaze calibration phase.

The gaze tracking module sends its estimation result to the interaction management module via socket communication.

[6]CPU: Core i7 2.93 GHz, Memory: 4 GB, OS: Windows 7

D. Content presentation module

The Info-concierge provides visual and audio information using our software based on Win32API. The screen, whose size is 1106 mm (1920 pixels) in height and 622 mm (1080 pixels) in width, is divided into four peripheral areas and an intermediate area. The system displays an article in each peripheral area (720 pixels in height and 405 pixels in width), and the anthropomorphic agent in the intermediate area. Each piece of content (each research topic) consists of five pages worth of articles.

We adopt a simple design using only the eyeballs in the agent's aspect. We suppose that the simple agent's aspect allows effective expression of the mirroring interaction. The system establishes joint attention by expressing the approach movement of the eyeballs toward each article. Although the agent would be able to express the gaze behavior by moving only the irises (without the approach movement), it is possible that the user will not be aware of Gaze Mirroring, because the visual variations are not large. As mentioned in the previous section, the gaze estimation accuracy is about 10 cm on the screen. To accurately estimate which article the user fixes his/her gaze on, each peripheral area is located at an interval of approximately 20 cm.

The system synthesizes the agent's speech explanation from text using the HOYA VoiceText engine SDK[7]. The text explaining each article contains about 100 characters in Japanese. It takes approximately 20 seconds to read.

The content presentation module is embedded in the interaction management module in the implementation and communicates with the content management module using a function-based library.

E. Content management module

The content management module receives gaze distribution on the screen (the degrees of interest) from the interaction management module and updates the distances between topics based on the distribution. The module then selects four topics (according to the distances) from a content database, that is, four topics similar to the most interesting topic on the screen.

The content management module is embedded in both the interaction management module and the content presentation module.

F. Interaction management module

The interaction management module has a timer to manage the flow of the interaction. This module receives keywords from the speech recognition module and eye movements from the gaze tracking module, and exchanges content and gaze distribution with content handle modules. The concrete behaviors of this module are described in Section III.

V. Field trial

This section presents a demonstration of Info-concierge and the parameters embedded in the demonstration system.

[7]Japanese female voice "SAYAKA"

Fig. 10. The scene of the field trial of Info-concierge.

We demonstrated the Info-concierge at the "New IT Infrastructures for Information Explosion Era (Info-plosion)" research project symposium on March 10 and 11, 2011 in Tokyo, Japan[8]. Fig.10 shows the scene of the demonstration. We also demonstrated another type of Info-concierge which provides information on social topics (ex. the electric cars are environmentally friendly?). Using this Info-concierge, users can understand their topics from various viewpoints and arrive at surprising information. We used the WISDOM system [20] that analyzes credibility of web information, to create the social content.

Approximately 30 people visited our site over the course of seven hours (without any breaks). Most of the visitors were researchers and had some discussions with us regarding the system while interacting with it. The average duration of interaction was about five minutes (excluding the duration of any discussion with the demonstrator). Some visitors could not have a smooth interaction with the system because of gaze tracking errors and ambient noises.

The system parameters of each phase were defined as follows:

- **Zapping support:** $T_{pause} = 3.0, \bar{T}_{enum} = 3.0, \hat{T}_{bargein} = 1.7, \sigma = 0.2$**,**
- **Gaze calibration:** $T_{interval} = 3.0, T_{gazeprobe} = 0.5$**,**
- **Awareness support:** $T_{mirror} = 15.0$**,**
- **Interest estimation:** $T_{engage} = 20.0, T_{interest} = 1.0$**,**
- **Recommendation:** $T_{recommend} = 10.0$**.**

We received high praise from many visitors, including from non-HCI researchers. They noted the effectiveness of both the sensing of the user gaze and the reaction timing. On the other hand, a number of visitors commented on the following need for improvements: the system should react to user utterances at any time, the system should contribute to a more varied dialogue, and the system should have access to the wealth of information available on the Internet.

[8]The symposium and the demonstration were closed early because of the effects of the earthquake disaster.

VI. Conclusions

We believe rich human-computer interaction contributes to a closing of the gap between the information world and the physical world. To achieve this goal, we have proposed a multi-modal interaction strategy called Mind Probing which consists of proactively acting on the user and estimating his/her internal state based on his/her reactions (without waiting for any commands from the user), and developed Info-concierge composed of interaction techniques based on Mind Probing. Info-concierge can probe the latent interest of the user and increase his/her awareness of it, and proactively provide sensible information. Through a field trial of Info-concierge using a large screen display, we confirmed the effectiveness of the techniques and their integration into Info-concierge. In the future, we will introduce Mind Probing into a facilitator system to support human-human communication and aim at the design of a novel communication environment.

Acknowledgments

This work was supported by the Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contract of numbers 18049046 and 23700168. We thank Prof. Y. Kuno (Saitama Univ., Japan), Prof. Y. I. Nakano (Seikei Univ., Japan), and Prof. K. Komatani (Nagoya Univ., Japan) for their contributions of research results for use with Info-concierge, and the students of Matsuyama Laboratory (Kyoto Univ., Japan) for their implementations of the demonstration system.

References

[1] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff, "Integration of Speech Recognition and Natural Language Processing in the MIT VOYAGER System," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1991)*, pp.713–716, 1991.

[2] R. S. Wallace, "The Anatomy of A.L.I.C.E," *Parsing the Turing Test*, Springer Netherlands, Part III, pp.181–210, 2009.

[3] R.W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.23, No.10, pp.1175–1191, 2001.

[4] T. Onishi, T. Hirayama, and T. Matsuyama, "What Does the Face-turning Action Imply in Consensus Building Communication?," *Proceedings of the 5th International Workshop on Machine Learning for Multimodal Interaction*, pp.26–37, 2008.

[5] P. Qvarfordt and S. Zhai, "Conversing with the User Based on Eye-gaze Patterns," Proceedings of the SIGCHI Conference on Human-Factors in Computing Systems, pp.221–230, 2005.

[6] E. André, T. Rist, and J. Müller, "Integrating Reactive and Scripted Behaviors in a Life-like Presentation Agent," *Proceedings of Agents*, pp.261–268, 1998.

[7] Y. Sumi and K. Mase, "AgentSalon: Facilitating Face-to-face Knowledge Exchange Through Conversations among Personal Agents," *Proceedings of Agents*, pp.393–400, 2001.

[8] K. Matsuyama, K. Komatani, T. Ogata, and H.G. Okuno, "Enabling a User to Specify an Item at Any Time During System Enumeration – Item Identification for Barge-In-Able Conversational Dialogue Systems," *INTERSPEECH2009*, pp.252–255, 2009.

[9] R. Yonetani, H. Kawashima, T. Hirayama, and T. Matsuyama, "Gaze Probing: Event-Based Estimation of Objects Being Focused On," *Proceedings of the 20th International Conference on Pattern Recognition*, pp.101–104, 2010.

[10] H.-S. Park, T. Hirayama, T. Matsuyama, "Gaze Mirroring-based Intelligent Information System for Making User's Interest Explicit," *Journal of Korea Intelligent Information Systems Society*, Vol.16, No.3, pp.37–54, 2010.

[11] Y. I. Nakano and R. Ishii, "Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations," *International Conference on Intelligent User Interfaces (IUI2010)*, pp.139–148, 2010.

[12] T. Hirayama, J. B. Dodane, H. Kawashima, and T. Matsuyama, "Estimates of User Interest Using Timing Structures between Proactive Content-display Updates and Eye Movements," *IEICE Transactions on Information and Systems*, Vol.E93-D, No.6, pp.1470–1478, 2010.

[13] Y. Kuno, K. Sadazuka, M. Kawashima, K. Yamazaki, A. Yamazaki, and H. Kuzuoka, "Museum Guide Robot Based on Sociological Interaction Analysis," *Proceedings of SIGCHI Conference on Human-Factors in Computing Systems*, pp.1191–1194, 2007.

[14] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero, "An Introduction to Voice Search," IEEE Signal Processing Magazine, May 2008.

[15] N. J. Emery, "The Eyes Have It: The Neuroethology, Function, and Evolution of Social Gaze," *Neuroscience and Biobehavioral Reviews*, No.24, pp.581–604, 2000.

[16] T. Misu and T. Kawahara, "Speech-based Interactive Information Guidance System Using Question-answering Technique," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol.IV, pp.145–148, 2007.

[17] A. Lee and T. Kawahara, "Recent Development of Open-source Speech Recognition Engine Julius," *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp.131–137, 2009.

[18] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *Proceedings of the 5th European Conference on Computer Vision*, Vol.2, pp.484–498, 1998.

[19] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle Adjustment – A Modern Synthesis," *Vision Algorithm: Theory & Practice, (B. Triggs, A. Zisserman, and R. Szeliski, eds.)*, Springer-Verlag LNCS 1883, 2000.

[20] S. Akamine, D. Kawahara, Y. Kato, T. Nakagawa, K. Inui, S. Kurohashi, and Y. Kidawara, "WISDOM: A Web Information Credibility Analysis System," *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pp.1–4, 2009.