# Multimodal Dialog System for Kyoto Sightseeing Guide

Hideki Kashioka, Teruhisa Misu, Etsuo Mizukami, Yoshinori Shiga,
Kentaro Kayama, Chiori Hori, and Hisashi Kawai
National Institute of Information and Communications Technology (NICT), Kyoto, Japan.
E-mail: hideki.kashioka@nict.go.jp

***Abstract*— We proposed a dialog system on Kyoto tourist information assistance in a client-server fashion. Our proposed system is called the "proactive dialog system" and aims to present acceptable information in an acceptable time. We developed two prototype systems. The first one is designed for mobile use. It was implemented in iPhone and its application is opened to the public in AppStore. The second one is designed for multi-modal information integration on large display panel. It can detect non-verbal information, such as changes in gaze and facial direction as well as head gestures of the user during dialog, and recommend suitable information. These two prototype client systems are basically connecting to the server module. This server module uses a weighted finite-state transducer (WFST) in which user concept and system action tags are input and output of the transducer. We implemented a dialog scenario to present sightseeing information on the system. In our proposed dialog system, we designed our system's behavior like human behavior. One of the most enduring problems in spoken dialogue systems research is realizing a natural dialogue in a human-human form. One-direction researchers have been utilizing spontaneous nonverbal and paralinguistic information. So that we collect human to human dialog corpus, and semi-automatically design a scenario which handles dialog in response to user' input so as to accomplish a task efficiently. Especially we focus on users' verbal feedback and non-verbal feedback in the form of nods. This paper presents our proposed system's outline and its function. After that in this paper, we display the results of an evaluation of image processing techniques for estimating facial direction from a camera for a multi-modal spoken dialog system on a large display panel. Experiments that consist of 100 sessions with 80 subjects were conducted to evaluate the system's efficiency. The system grows particularly clear when dialog contains recommendations.**

## I. INTRODUCTION

Nowadays we can get most information through the Internet. However, we have a trouble to pick up expected information from the huge results with conventional search engines. Then we are confronted with great difficulties for two factors. One is that most of users cannot make an appropriate query because their request is vague with themselves. The other is that the retrieved information has huge variation and client terminal has not enough area for displaying them. Therefore, we aim to develop technologies for the users to input their requests by familiar way and clarify what they want to know with displaying the retrieved information with suitable method.

Our development system is baseline system providing Kyoto tourist information in a client-server fashion [1]. And this system is constructed with two major parts (Fig. 1). One is a dialog management part as dialog server, and the other is information analysis part as knowledge access server. Dialog management part mainly includes following modules: speech recognition, dialog control, and speech synthesis. Additionally image-processing module would be used for recommend suitable information. Information analysis part has two major functions. One is ``Reputation Search" function that is realized in WISDOM (Web Information Sensibly and Discreetly Ordered and Marshaled)[2]. And the other is ``Associative Keyword Search" function that is realized in ``knowledge cluster system"[3].

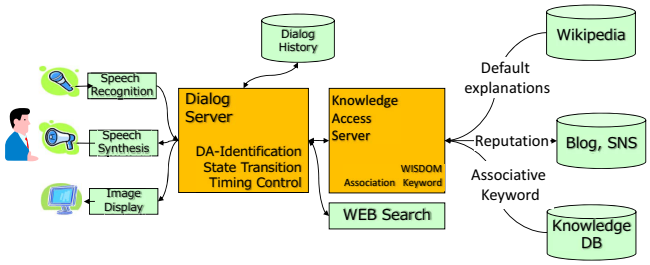


**Fig. 1 Basic Module Construction for Spoken Dialog System**

Image processing techniques estimating face and gaze direction from cameras have been widely studied in recent years, and these techniques are used as a multimodal user interface because the directions are thought to indicate the user's attention [2,3]. It is, however, difficult to evaluate the efficiency of these techniques in multimodal applications because many factors influence user's impressions. In this paper, we describe a client module with multi-modal information integration on a large display panel. This system would be used for digital signage using image-processing techniques, evaluated the performance of image processing and its efficiency in our application based on dialog corpora and videos of 100 sessions from which 80 subjects actual spoke to our system. "Digital Signage" is an advertising media for selecting and displaying in real time appropriate contents according to the users, and has been actively studied in recent years [4]. Almost all of these, however, one-sidedly display content, or need explicit input devices such as touch panels. We expect natural interfaces make these systems friendlier for user. For example, the ability for a user to draw out desirable information from a system via spoken dialogs and for the system to predict the user's interests and recommend appropriate information would produce an ambient intelligence. The construction of such systems can lead to applications such as next-generation digital signage. Therefore we proposed a novel interactive information display system that realizes proactive dialogs between human

and computer based on image processing techniques [5,6]. A "proactive dialog system" refers to a system that has the functionality of actively presenting acceptable information in an acceptable time in addition to being able to adequately respond to queries [7]. The proposed system that integrated image-processing information in a client module is based on spoken dialog. It is also able to detect non-verbal information, such as changes in gaze and facial direction and head gestures of the user during dialog, and recommend appropriate information. We constructed the prototype of this system with data and dialog scenarios for sightseeing guidance on Kyoto. Experiments were held with 80 subjects (total 100 sessions) to analyze user behavior during system use and to evaluate the system's usefulness and performance of image processing used in the system. In this paper, we present software and hardware architecture, image processing technology for the system, and user evaluations. Hardware and software architecture and details of the parts of spoken language recognition and display control are described in section 2. Image processing that detects users and estimates gaze and facial directions is explained in section 3. The application implemented in this system is described in section 4. User's spontaneous backchannel is reported in section 5.


**Figure 1** **Screen example of PDP information dialog system**

## II. SYSTEM ARCHITECTURE

### *A. Outline of Total System*

Dialog systems help users accomplish a task through (spoken or multi-modal) interaction with machines [8]. When a user requests something by speaking and/or pointing to a system, the system responds to the user's request. Therefore, to design a complex scenario which absorbs users' spontaneity, we may need to combine several scenarios written in different fashions such as finite-state automaton, frame-based representation, if-then rules, etc. We can also take stochastic approaches such as partially observable Markov decision process (POMDP) [9] when the model can be trained appropriately with enough data. However, it is not easy to control different fashions of scenarios in a dialog system considering multi-modal inputs, and therefore enormous labor is required to implement such functions. Accordingly, we need an expandable and adaptable integration platform that enables us to separately design several scenarios and functions to handle system actions in response to user's input. In Figure 2 we design dialog management WFST constructed with four different WFSTs; 1) main scenario WFST, 2) task dependent scenario WFST, 3) task independent WFST, 4) spoken language understanding (SLU). Each WFST are manually or semi-automatically constructed from human to human dialog corpus.

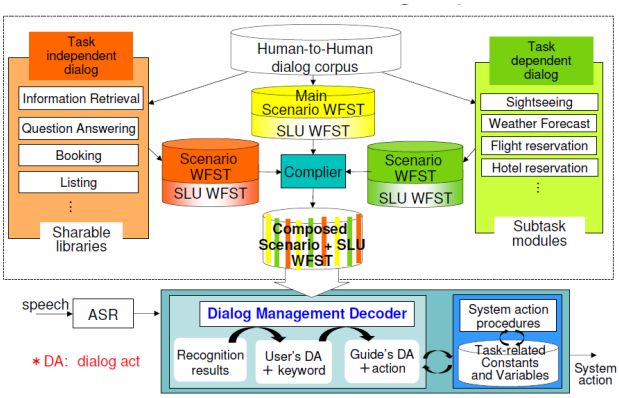


**Figure 2** **WFST-based Dialog System**

This section discusses one of the systems that we proposed prototype of spoken dialog systems. This system is used with a plasma display panel that we constructed for a system integrating non-verbal information recognition and spoken dialog. The system is constructed on the premise of it being fixed in a public space, such as a tourist information office, and presenting information to a general audience. The main input interface is also assumed to be spoken language and image processing is used to enhance dialog quality by estimating user interests. The output interface utilizes a wide screen divided into four windows and displaying a range of information. The character shown in Figure 1 appears on the screen and explains displayed content and controls dialog via speech synthesis.

### B. *Hardware*

The prototype of the proposed system we constructed is shown in Figure 3.

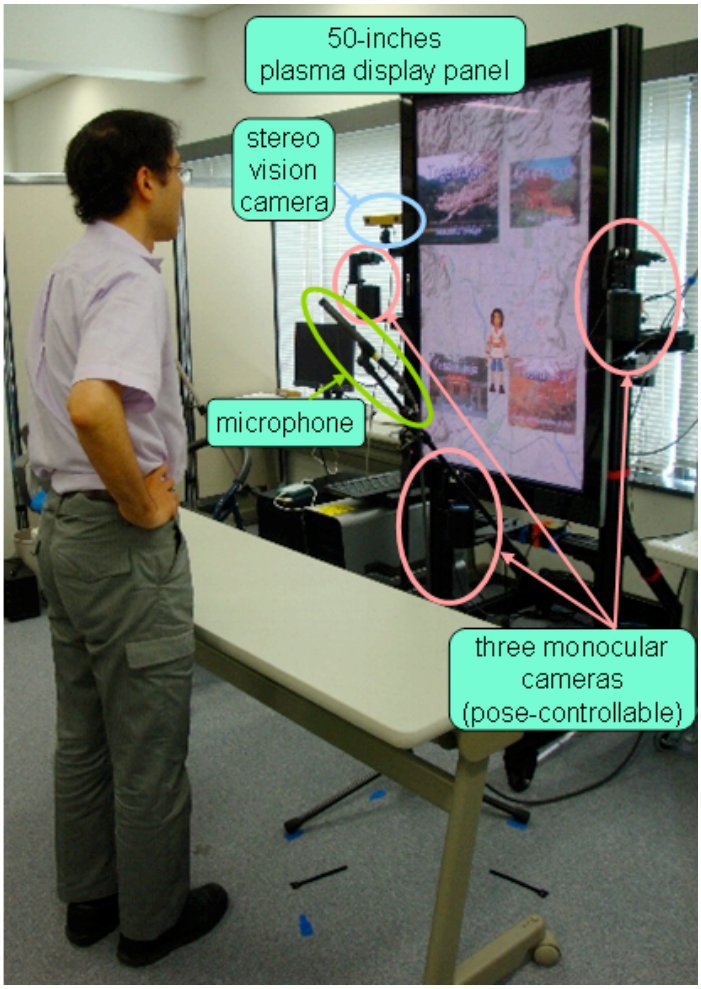


**Figure 3 Spoken dialog system on plasma display panel**

It consists of following parts:

- **50-inch plasma display panel (PDP)**
  70-cm wide and 120-cm high portrait display on an 80-cm-high base with a resolution of 1080 X 1920.
- **Three pose-controllable monocular cameras**
  Grasshopper made by Point Grey Research Inc. with attached lenses so that face of a user is correctly placed in view.
- **Stereo vision camera.**
  Bumblebee2 also made by Point Gray Research Inc. , with a horizontal angle of about 60-degree width.
- **Directional microphone.**
  CS-3e made by SANKEN
- **Loudspeaker.**
  PN-AZ10 made by Sony
- **9 PC**
  5 PC for image processing, 2 PC for speech input /recognition processing, 1 PC for display control, 1 PC for speech synthesis

### C. *Software*

Software Modules are divided into these four function types:

1. **Image processing**
   The image processing algorithms implemented in the system are human and head area detection and face and gaze direction estimation. The details of the algorithms are described in the next section
2. **Speech recognition and parsing**
   First, voice activity detection (VAD) is performed for the input audio signal and the uttering part is cut out. This part is sent to a module that contains ATRASR [10] which was developed as a speech recognition engine and performs speech recognition and parsing. Parsed results are sent finally to the dialog control.
3. **Dialog control**
   The basic dialog control mechanism is described in outline for total system. In this system, we adopt WFST-based mechanism for dialog control.
4. **Display control and speech synthesis**
   The display control outputs a screen as in Figure 1. In principle, the screen is divided into two or four windows; with each window respectively using HTML. In Figure 1, a brief on Kinkaku-ji temple is displayed on the upper left, a list of restaurants near Kinkaku-ji on the lower left, details of a restaurant on the lower right. The character agent is displayed on the center of the lower right window. She performs several actions as a virtual conversational partner. She is equipped with lip synchronization. The shape of her mouth is generated based on the vowel by cooperating with the speech synthesis.

## III. NON-VERBAL INFORMATION PROCESSING VIA IMAGES

### A. *Detection of Head Area*

Candidates for the human head area via use of the stereo vision camera are detected with the following processes:

1. Construct three-dimensional occupancy grid (10-cm size)
2. Divide into each 20-cm height and cluster object existence area
3. Segment to each person area via an applied method of the crossing hierarchy method [11]
4. Detect candidate areas of the human head
5. Evaluate and filter candidates by evaluating head possibility

An example of the processing is shown in Figure 4 In the crossing hierarchy method, three-dimensional space is divided into overlapping plates: for example, 20-cm plates at heights of 180–160 cm, 170–150 cm, 160–140cm. Areas that are determined as the human region in the upper unit are then propagated to lower units and candidate areas of the human region are decided sequentially. In this way entire human regions are abstracted. The original method used multiple stereo vision cameras, but this system uses only one.

Moreover, the camera of our system is equipped in plan-view, which causes serious occlusion but makes setting easy.

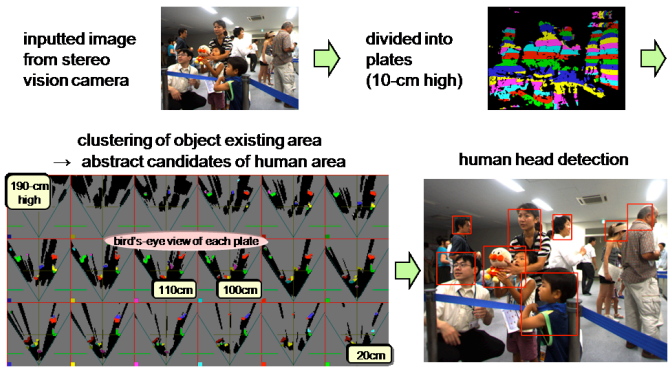


**Figure 4 Person detection from stereo images**

In this system, clustering is first performed on each plate. A distinction is then made for each cluster based on whether it will inherit a cluster of the upper plate or appear as a new region. If a cluster inherits multiple regions it is adequately divided, and if multiple clusters inherit a single region they are integrated, if necessary. These processes enable robust detection of each human region when there are occlusions, closes, and contacts with multiple people. The center part of Figure 4 shows 18 plates at heights ranging from 200–180 cm to 30–10 cm, used in this system. PDP exists in the lower center of each square. No objects are black regions. Grey regions indicate invisible areas that are out of view or occluded. Other colored regions are candidates for the human region. The same color indicates the same person.

After that, the upper region, which is about 30 cm from the top of the human region, is detected as a candidate area for human head. Moreover, the possibility that each candidate area is a human head is evaluated on its height, distance from the PDP, size, shape, and head position of the previous frame. Areas that exceed a priori threshold are finally decided to be the human head.

### *B. Facial Direction Estimation*

The system controls three high-resolution monocular cameras to catch human head regions obtained from the above mentioned processes. Facial direction is then estimated as follows by individual monocular cameras (Figure 3), and mean of the cameras are used to know where a user looks at:

1. **Detect face regions.**
   In this system, images of 800 x 600 pixels are input at 15 frames per second. If the system fails to detect a face or to track facial parts in the image, in the next frame a facial detection routine using Haar-like features is executed.
2. **Detect and track facial parts.**
   The system detects 45 feature points on the face by using an active appearance model (AAM), as initial values of the coordination of points are the values of the previous frame (if facial parts were also detected in the previous frame) or a priori values (if the face is newly detected) [12]. AAM is a method for using principal component analysis against the vector consisting of coordination of the features of facial parts in the image and intensities of the pixels of the face region. The correlation in the change of the feature point locations and the change of view is learned. It enables tracking non-rigid objects such as facial parts.
3. **Estimate facial direction.**
   Six degrees of freedom (DOF), i.e. three in rotation and three in translation, of the facial direction are estimated by using the steepest descent method by fitting three-dimensional coordination of each feature in an a priori three-dimensional face-shape model and the calculated coordination of the previous step.
4. **Estimate gaze direction.**
   The candidate for the iris region is obtained by binary eye regions and fitting an ellipse. Moreover, three-dimensional coordination of the center of the eyeball, obtained in the previous step, and coordination of the center of the iris calculated via the coordination of the candidate of the iris region in the image and facial direction are estimated. Gaze direction is the direction of the line on these two points [13].

We use facial direction to know where a user is looking because iris detection is not robust given the variety of individual eye-shape, and lighting noise in the real world. We define a baseline of facial direction from a center of the PDP to a facial gravity-point of a user who is looking at the center of PDP. Blue lines in Figure 5 show the base lines. This system calculates the motion of the base-line from the Six DOF. Finally this system identifies the point where the baseline crosses the PDP plane.

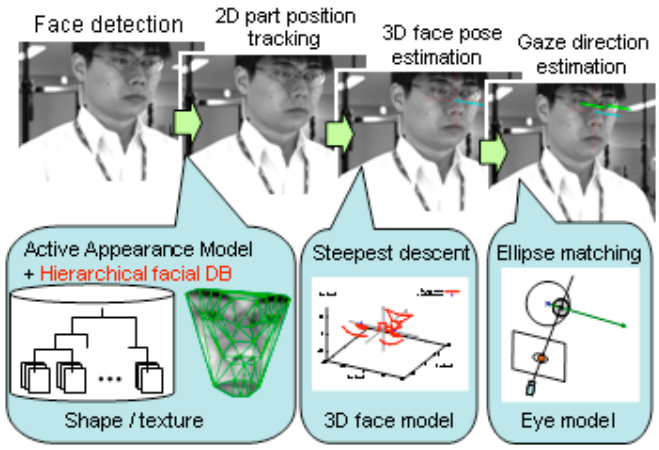


Figure 5 Facial direction estimation

## IV. APPLICATION

### A. Acceptable Queries

We implemented a Kyoto sightseeing guide scenario on the prototype system, based on a system for a portable PC [14]. Examples of acceptable queries on the system are as follows:

1. "Show me sightseeing spots that are famous for determinant. ".
2. "Show me sightseeing spot. ".
3. "Show me subject of sightseeing spot. ".
4. (If search result list is displayed) "Show me details on the n-th item."

We assigned cherry blossoms, autumn foliage, and gardens as determinant. A bus schedule, how to go to the site, a map, restaurants near the site, and so on are prepared as a subject in 3. When words are recognized, the system displays the contents. If a sightseeing spot is omitted, the system infers that the spot mentioned just before is also a current sightseeing spot. The system has a database of about 2,000 sites. If a user utters words not contained in the databases, the system searches by treating the words as keywords on Google and displays the results in list form. If a user utters a number of such as "4," the website nominated by the user is displayed.

### B. Recommendation Based on Non-verbal Information

A typical function of this dialog system is recommendation dialogue based on non-verbal information. This system shows information in four or two windows on the PDP as shown in Figure 1, and this system automatically recommends contents by estimating which window the user is looking at, when the user doesn't know what to say 1. Finally this system transits the state automatically with a system utterance, "I will explain this content." Figure 6 shows details of the dialog status.

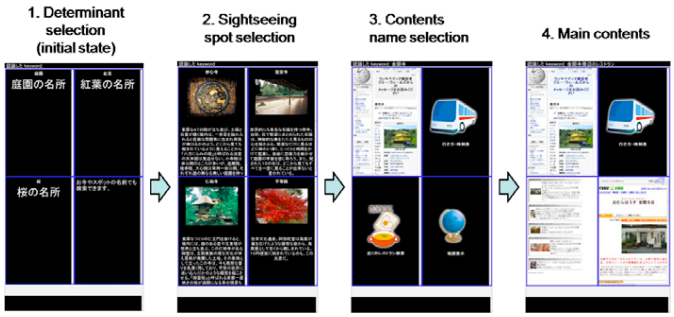


**Figure 6 Dialog state transition**

1. Initial state: Four determinants are displayed. "Determinants" are displayed on the four-section screen.
2. Four sightseeing spots are displayed. In this mode, pictures and a brief explanation of four spots that fulfill the condition selected in step 1 are randomly selected and displayed.
3. Four content names are displayed. In this mode, an abstract of sightseeing spots selected in step 2 is displayed in the upper left of the screen. In addition, the types of information searchable in the system, such as "how to go to the site," "map," and "restaurant near the site" are displayed.
4. Two or four contents are displayed. Contents mean a map near the site of the current topic, list of restaurants near the site, details of one restaurant on the list, bus schedule from Kyoto station to the spot, and so on. When the system estimates that a user utters a query that requests content that is not prepared, the result list from a Google search for the keyword contained in the utterance and the first-ranked web-site on the list is displayed.

These states are supposed to transit from 1 to 2, 2 to 3, and 3 to 4 as a standard. Regardless of the current state, however, uttering of "determinants" causes transition to state 2, uttering something such as "Show me sightseeing spot" (sightseeing spot is different from current topic) causes transition to state 3, and uttering something such as "Show me subject of sightseeing spot " to state 4. If a human is not detected by the image processing for a certain period or the user says "Thank you," the system is reset and returns to state 1.

## V. USERS' VERBAL FEEDBACK AND NON-VERBAL FEEDBACK

In this section, we considered with para-linguistic information in human-human dialogue. We focus on users' verbal feedback, such as "uh-huh" (called Aizuchi in Japanese), and non-verbal feedback in the form of nods. These phenomena is one of the most common forms as called backchannels, and considered to be used to facilitate smooth human-human communications. In this regard, Maynard [15] indicated that such backchannels are listener's signals to let the speaker continue speaking (continuer), to indicate that the listener understands and consents. It was also hypothesized that humans detect feelings expressed via backchannels, and the correlation between backchannel patterns and user interests was examined [16]. These studies indicate that detection of spontaneous user backchannels can benefit spoken dialogue systems by providing informative cues that reflect the user's situation. For instance, if a spoken dialogue system can detect user's backchannels, it can facilitate smooth turn taking. The system can also detect user's feelings and judge if it should continue the current topic or change it.

Despite these previous studies and decades of analysis on backchannels, few practical dialogue systems have made use of them. This is probably due to the fact that users do not react as spontaneously to dialogue systems as they do to other humans. We presume one of the reasons for this is the unnatural intonation of synthesized speech. That is, conventional speech synthesizers do not provide users with signs to elicit backchannels; an appropriate set of lexical, acoustic and prosodic cues (or backchannel-inviting cues [17]), which tends to precede the listener's backchannels in human-human communication. Though recorded human speech can provide such cues, it is costly to re-record system's speech every time system scripts are updated.

Then we develop our dialog system how to treat these matter under the hypothesis of :

*People will give more spontaneous backchannels to a spoken dialogue system that makes more spontaneous backchannel-inviting cues than a spoken dialogue system that makes less spontaneous ones.*

which is derived from the Media Equation [18].

So we construct a spoken dialogue-style HMMbased TTS (text-to-speech) system and then analyze if the synthesized speech has backchannel-inviting prosodic cues. The TTS is evaluated in terms of user's evoked listener's reactions. We demonstrate that our TTS system can invite more spontaneous user backchannels than the conventional one. Our experimental results suggest that our dialogue-style TTS system can evoke more spontaneous and informative backchannels that reflects users' intentions than the conventional reading-style one. This classification rate is not completely satisfactory, but we expect that users' feeling can be detected after observing several backchannels. We also believe that we can estimate users' interest more precisely by combining verbal information of dialogue acts [19].

## CONCLUSIONS

We constructed a prototype of a proactive spoken dialog system, which is a smart interactive information presentation system that incorporates a non-verbal information recognition process into a spoken dialog system. The system consists of a 50-inch plasma display panel, microphone for voice input, and cameras (three monocular and one stereo) for image input.

We implemented a Kyoto sightseeing guide scenario on the system, in which dialog mainly progresses by spoken language. When there is no utterance for a given duration, the system recommends information by estimating user interests via facial direction estimated by image information.

We conducted an experiment on the system with 100 sessions to evaluate its efficiency. The head detection rate was almost 100%, but the success rate of detecting where a user was looking was about 30%. This seemed to be the reason why the recommendation acceptance rate was about 30%. Evaluations for the system with recommendations, however, were mostly higher than those without recommendations. In particular, the evaluation of the system's clarity increases when dialog provides recommendations.

We constructed a dialogue-style TTS and confirmed that by generating human-like backchannel-inviting cues, the system can evoke user's spontaneous backchannels, which are informative for the system. Our user experiment focused only on prosodic features, and analyses and experiments are needed before clarifying the mechanism of users making backchannels.

## REFERENCES


[1] S. Nakamurai, Spoken Language Technologies for Universal Communication, Proc. of the First International Symposium on Universal Communication, 2007.

[2] Kobayashi, Y., Sugimura, D., Sato, Y., Hirasawa, K., Suzuki, N., Kage, H., Sugimoto, A.: 3D Head Tracking using the Particle Filter with Cascaded Classifiers. In: Proc. British Machine Vision Conference (BMVC 2006), pp. 37–46 (2006)

[3] Fujie, S., Yamahata, T., Kobayashi, T.: Conversation robot with the function of gaze recognition. In: Proc. 2006 IEEE-RAS Int'l Conf. on Humanoid Robots (Humanoids 2006), pp. 364–369 (2006)

[4] Lao, S., Yamaguchi, O.: Facial Image Processing Technology for Real Applications: Recent Progress in Facial Image Processing Technology. IPSJ Magazine 50(4), 319–326 (2009)

[5] Kobayashi, A., Kayama, K., Lee, D., Sumi, K., Kato, T., Kadobayashi, R., Yamazaki, T.: Proposition of Proactive Information Display System Using Face Directions and Head Gestures Estimation. In: Proc. of the IEICE General Conference (2009)

[6] Minakuchi, M., Asano, S., Satake, J., Kobayashi, A., Hirayama, T., Kawashima, H., Kojima, H., Matsuyama, T.: Mind Probing: Active Stimulation of Gaze Patterns for Inference of User's Interest. Information Processing Society of Japan (IPSJ) SIG Technical Reports (Human-Computer Interaction, HCI) 125, 1–8 (2007)

[7] Kawahara, T., Kawashima, H., Hirayama, T., Matsuyama, T.: "Automated Information Concierge" based on Proactive Dialog and Information Retrieval. IPSJ Magazine 49(8), 912–918 (2008)

[8] J. Glass, et al., "Facilitating spoken dialogue system development," In Proc. Interspeech2001.

[9] J. Williams and S. Young, "Partially Observable Markov Decision Processes for Spoken Dialog Systems." Computer Speech and Language 21(2): 231-422, 2007.

[10] Itoh, G., Ashikari, Y., Jitsuhiro, T., Nakamura, T.: Summary and evaluation of speech recognition integrated environment ATRASR. In: Proc. of 2005 Acoustic Society of Japan Fall Meeting, pp. 221–222 (2005)

[11] Yoda, I., Sakaue, K.: Concept of Ubiquitous Stereo Vision and Applications for Human Sensing. In: Proc. on, 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA 2003, pp. 1251–1257 (2003)

[12] Kobayashi, A., Satake, J., Hirayama, T., Kawashima, H., Matsuyama, T.: Person-Independent Face Tracking Based on Dynamic AAM Selection. In: IEEE Int. Conf. on Automatic Face and Gesture Recognition, FG (2008)

[13] Satake, J., Kobayashi, A., Hirayama, T., Kawashima, H., Matsuyama, T.: Accuracy Improvement of Real-Time Gaze Estimation using High Resolution Camera, Technical report of IEICE. PRMU 107(491), 137–142 (2008)

[14] Kashioka, H., Misu, T., Ohtake, K., Hori, C., Nakamura, S.: Development of dialog system keeping step with users, Technical report of IPSJ. SLP 2008(68), 93–97 (2008)

[15] S. Maynard.: On back-channel behavior in japanese and english casual conversation. Linguistics , 24(6):1079–1108. 1986.

[16] T. Kawahara, M. Toyokura, T. Misu, and C. Hori.: Detection of Feeling Through Back-Channels in Spoken Dialogue. In Proc. Interspeech , pages 1696–1696. 2008.

[17] A. Gravano and J. Hirschberg.: Backchannelinviting cues in task-oriented dialogue. In Proc. Interspeech , pages 1019–1022. 2009.

[18] B. Reeves and C. Nass.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places . Cambridge University Press. 1996.

[19] Teruhisa Misu, Komei Sugiura, Kiyonori Ohtake, Chiori Hori, Hideki Kashioka, Hisashi Kawai, and Satoshi Nakamura. Dialogue Strategy Optimization to Assist User's Decision for Spoken Consulting Dialogue Systems. In Proc. IEEE-SLT , pages 342–347. 2010.