# Eye Movement as an Indicator of Users' Involvement with Embodied Interfaces at the Low Level

Chunling Ma[⋆]
[⋆]Dept. of Information and Communication Eng.
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
macl@miv.t.u-tokyo.ac.jp

Helmut Prendinger[†]
[†]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
helmut@nii.ac.jp

Mitsuru Ishizuka[⋆]
[‡]Dept. of Information and Communication Eng.
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
ishizuka@miv.t.u-tokyo.ac.jp

**Abstract**

In this paper, we motivate an approach to evaluating the utility of animated interface agents that is based on human eye movements rather than questionnaires. An eye tracker is employed to obtain quantitative evidence of a user's focus of attention. The salient feature of our evaluation strategy is that it allows us to measure important properties of the user's interaction experience on a moment-by-moment basis. We describe an empirical study in which we compare attending behavior of participants watching the presentation of an apartment by three types of media: an animated agent, a text box, and speech only. Users' eye movements may also shed light on their *involvement* in following a presentation.

## 1 Introduction

Animated interface agents have attracted considerable interest and attention in recent years, mainly for their ability to emulate human–human communication styles that is expected to improve the intuitiveness and effectiveness of user interfaces (see, e.g. André et al. (1996) for early work in this area). Following this user interface paradigm, a considerable number of animated agent (or character) based systems have been developed, ranging from information presentation and online sales to personal assistance, entertainment, and tutoring (Cassell et al., 2000; Prendinger and Ishizuka, 2004b). While significant progress has been made in individual aspects of animated agents, such as their graphical appearance or quality of synthetic voice, evidence of their positive impact on human–computer interaction is still rare. The most well-known evaluation studies have been directed towards showing the 'persona effect', stating that animated agents can have a positive effect on the dimensions of motivation, entertainment, and perceived task difficulty (Lester et al., 1997; van Mulken et al., 1998).

A common feature of most evaluations of interface agents is that they are based on questionnaires and focus on the user's experience with the systems hosting them, including questions about their believability, likeability, engagingness, utility, and ability to attract attention. However, as Dehn and van Mulken (2000) pointed out, the broad variety of realizations of animated agents and interaction scenarios complicates their comparison. More importantly, subtle aspects of the interaction, such as whether users pay attention to the agent or not, cannot be deduced reliably from self-reports (Nisbett and Wilson, 1977).

Furthermore, the concept of a user's *involvement* when interacting with computers has recently been discussed in the areas of Social Intelligence Design and Conversational Informatics (Nishida, 2001; Prendinger and Ishizuka, 2004a). While the term "involvement" embraces a wide range of concepts, including immersion and engagement (over a possibly extended time span), we want to consider involvement *at the low level* – whether the user is attending to designated objects of the interface – which seems to be an important prerequisite for any 'higher level' notion of involvement.

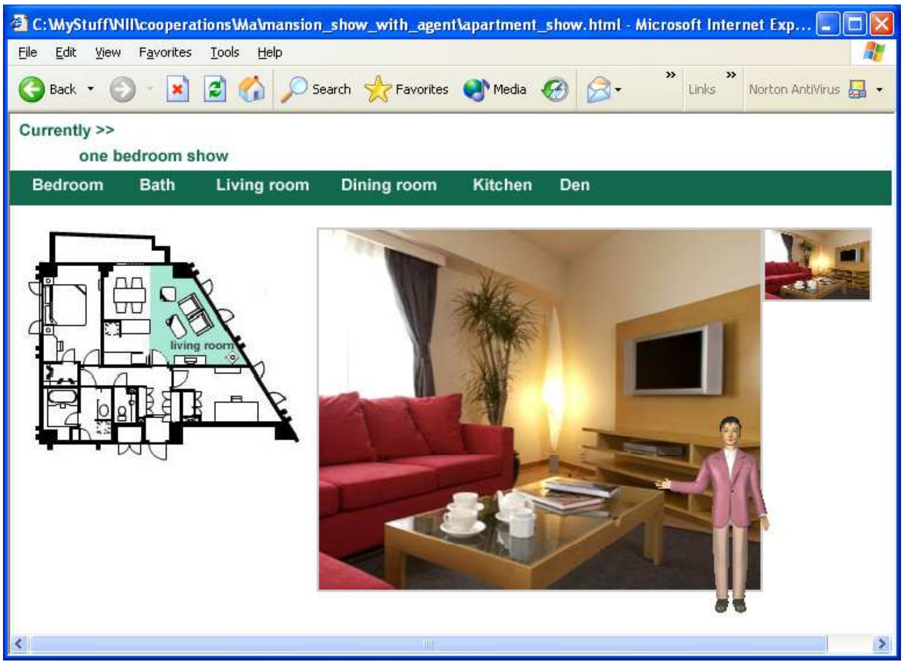


Figure 1: An animated agent presents the living room.

In this paper, we want to propose a different approach to evaluating animated agents, one that is based on eye movement behavior of users interacting with the interface. Although gaze point and focus of attention are not necessarily always identical, a user's eye movement data provide rich evidence of the user's visual and (overt) attentional processes (Duchowski, 2003). The movements of the human eye can be used to answer questions such as:

- Is the user paying attention to the interface agent?
- To which part of the agent (face or body) is the user attending to?
- Can the agent's verbal or gestural behavior direct the user's focus of attention?

Hence, eye movement data can offer valuable information relevant to the utility of animated agents and the usability of interfaces employing those agents. The tracking of eye movements lends itself to reliably capturing the moment-to-moment experience of interface users, which is hard to assess by using post-experiment questionnaires.

We will track and analyze eye movements while users are following the web page based presentation of different rooms of an apartment. Three types of presentations will be contrasted: (i) An animated interface agent presents the apartment using speech and gestures (see Fig. 1), (ii) the apartment is presented by means of a text-box and read out by speech, or (iii) the presentation is given by speech only.

The rest of the paper is organized as follows. The next section overviews work related to using eye movement as an evaluation method for user interfaces. The core part of the paper (Sect. 3) is devoted to the description of an experiment that provides both spatial and temporal analyses of users' eye movements during a presentation. The paper is rounded off by conclusions.

## 2 Related Work

This section reports on work that employs eye movements in the context of user interfaces. Eye movement data have been analyzed for two purposes, *diagnostic* and *interactive*. In the diagnostic use, eye movement data provide evidence of the user's attention and can be investigated to evaluate the usability of interfaces (Faraday and Sutcliffe, 1996; Goldberg and Kotval, 1999). In the interactive use, a system responds to the observed eye movements and can thus be seen as an input modality (Duchowski, 2003).

Goldberg and Kotval (1999) performed an analysis of eye movements in order to assess the usability of an interface for a simple drawing tool. Comparing a 'good' interface with well-organized tool buttons to a 'poor' interface with a randomly organized set of tool buttons, the authors could show that the good interface resulted in shorter scan paths that cover smaller areas. The measure of interest in their study is efficient scanning behavior, i.e. a short scan path to the target object. While this measure might not have high priority in our application domain, the merit of this study is to have introduced a systematic classification of different measures based on (temporal) scan paths rather than on cumulative (spatial) fixation areas. The temporal succession of transitions between different areas of attention is particularly relevant to investigate the effect of deictic references of animated agents to interface objects.

Faraday and Sutcliffe (1996) investigated attentional processing and comprehension of multimedia presentations. Core findings of the authors relevant to our domain can be summarized along the following dimensions:

**Shifts of attention.**

- A moving interface object induces a shift of attention to the object in motion.
- Attention is re-oriented when the presentation scene shifts.
- Labelling a presentation object produces fixation shifts between the object and the label.

**Locked attention.** A viewer's attention is locked when a moving object is processed, so that other presentation objects which are concurrently changed are not attended to.

**Auditory language processing and attention.** Comprehension of objects being presented visually with a spoken comment is increased only if both media types produce a single unified proposition.

The last mentioned item has also been investigated by Cooper (1974) who reports that people who simultaneously listen to speech and a visual object featuring elements that are semantically related to the spoken information tend to focus on the elements that are most closely related to the meaning of the currently heard spoken language (see also Duchowski (2003, p. 167)).

Witkowski et al. (2001) employ eye-tracking technology in order to assess user attention while interacting with an animated interface agent based online sales kiosk. In this setting, the interface agent provides help to the user and presents a product (a selection of wines). The authors conjecture that the agent will direct the attention of the users to the item of interest (help buttons, pictures of wines), following the agent's verbal comments. However, the results of their study do not support this hypothesis. In the experiment, a character agent controlled by the Microsoft Agent package (Microsoft, 1998) has been chosen with the text balloon enabled that depicts the text that is currently being spoken. The results reveal that users mostly focus on reading the text, rather than attending to the agent or to the product. In our study, we thus decided to disable the text balloon in order to avoid this problem. For the time that users were looking at the agent (face, gesture, body), the face was focussed on the most. In general, Witkowski et al. (2001) observed that interface agents do attract the attention of users. Similar results have been obtained by Takeuchi and Naito (1995) who compared an interface featuring either a (facial) agent or an arrow.

Besides its diagnostic role, eye movement data have also been used as an additional input modality to character-based intelligent systems. For instance, Qu et al. (2004) consider a user's focus of attention (among others) to decide an appropriate response in the context of educational software, and Nakano et al. (2003) investigate attentional focus (among others) in the setting of a direction-giving task.

# 3 Method

## 3.1 Experimental Design

A presentation of an apartment located in Tokyo has been prepared using a web page based interface (Mansions, 2004). The apartment consists of six rooms: living room, bedroom, dining room, den, kitchen, and bathroom. Views of each room are shown during the presentation, including pictures of some part of the room and close-up pictures of e.g. a door handle or sofa. Three versions of the apartment show have been designed for the experiment:

- *Agent (& speech) version.* A character called "Kosaku" presents the apartment using synthetic speech and deictic gestures (see Fig. 1). The character is controlled by a version of MPML (Prendinger et al., 2004).
- *Text (& speech) version.* The presentation content of each scene is displayed by a text box and read out by Microsoft Reader.
- *Voice (only) version.* Synthetic speech is the only medium used to comment on the apartment.

The main purpose of programming the Text and Voice versions was to provide interfaces that represent conceivable presentation types and can be compared to the Agent version in terms of the user's eye movements. The same type and speed of (synthetic) voice was used in all versions.

### 3.2 Subjects

Fifteen subjects (3 female, 12 male), all students or staff from the University of Tokyo, participated in the study (5 in each version). Their age ranged from 24 to 33 (mean 28.75 years). They were recruited through flyers and received 1,000 Yen for participation. In some cases the calibration process of the eye tracker was not successful due to reflections of contact lenses. Those subjects were excluded from the experiment beforehand.

### 3.3 Apparatus

The presentation of the apartment was hosted on a computer with a 17 inch (42.5 cm) monitor (the main monitor). A second computer was used to control the eye tracking system, a NAC Image Technology Eyemark Recorder model EMR-8B (NAC, 2004). The eye mark recorder is shown in Fig. 2 and the experimental setup is shown in Fig. 3.


Figure 2: NAC EMR-8B eye tracker.

The EMR eye tracker uses two cameras directed toward the subject's left and right eye, respectively, to detect their movements by simultaneously measuring the center of the pupil and the position of the reflection image of the IR LED on the cornea. A third camera is faced outwards, in the direction of the subject's visual field, including the main monitor. The system has a sampling rate of 60 Hz. The subject's head posture was maintained with a chin rest, with the eyes at a distance of 24 inch (60 cm) from the main monitor. A digital video recorder that captured the data from the third camera was connected to the computer that processed the eye movements.

The subjects were also connected to a bio-signal encoder that provides skin conductance (SC) and heart rate (HR) sensors. The bio-signal part of the experiment will not be reported here.

### 3.4 Procedure

Subjects were first briefed about the experiment. They were told that an apartment will be shown to them, and that they would be asked general questions about the apartment afterwards. They were also instructed to watch the demonstration carefully since they should be able to report features of the apartment to others.

The subjects were first put on the cap with the eye tracker. Calibration was performed by instructing the subject to fixate nine points in the screen area. After that, the subjects were shown the presentation that lasted for 8 minutes. Finally, the subjects were freed from the tracking equipment, and asked to fill out a questionnaire in order to report on their perception of the interface and to answer some content-related questions concerning the presented material.

### 3.5 Data Analysis

For analysis, the recorded video data of a presentation were first divided into individual scenes. A scene is a presentation unit where a referring entity (agent, text box, or voice) describes a reference object (an item of the room). Only the Agent and Text versions feature a visible referring entity. In Fig. 1, the scene consists of the agent performing a hand gesture to its right and introducing the living room. In order to be able to compare the three versions, scenes where the agent or text box moves from one location were left out. For each scene (41 in total), the following four screen area categories were defined:

- The area of a (visible) referring entity is either the smallest rectangle demarcating the agent or the text box (the agent area is further subdivided into face and body areas).
- The area of the reference object is the smallest rectangle demarcating the object currently described.
- The layout area (a designated, permanent reference object) is the field on the screen that displays the layout of the room.
- Other screen areas.

A program has been written that first maps eye-tracking data to $xy$-coordinates of the video sequence, and then counts the gaze points in each of the four categories.

When eye movements are relatively steady for a short period in one area, they are called *fixations* whereas rapid shifts from one area to another are called *saccades* (Duchowski, 2003). During a saccade, no visual processing takes place. If a cluster of

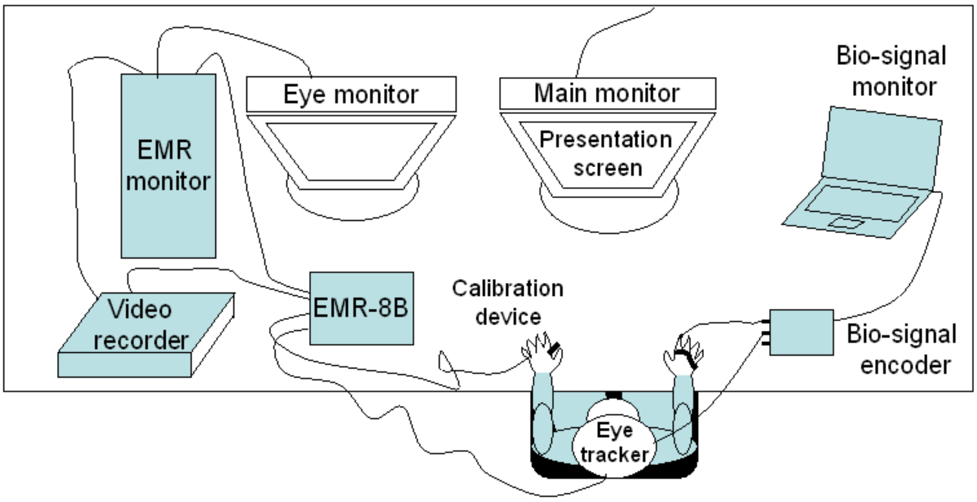


Figure 3: Experimental setup.

gaze points has less than 6 entries, it was categorized as part of a saccade (Goldberg and Kotval, 1999). All data accounted for in the analysis are derived from the activity of subjects' left eyes.

### 3.6 Results

The core of our results was distilled from analyzing subjects' eye movements. The level of statistical significance was set to 5%.

**Focus of Attention Hypothesis.** The ability of the interface to direct a subject's focus of attention to reference objects has been tested in two ways, spatial and spatio-temporal. The *spatial* analysis counts the gaze points that fall within areas of interest, specifically the reference object area and the layout area. Except for the introductory episode, the layout is not explicitly referred to during the presentation although it may serve as an orientation aid for users. The hypothesis is tested by restriction to those scenes where the referring entity (agent, text, voice) refers to some item of the apartment. An between–subjects analysis of variance (ANOVA) showed that users focus on the reference objects more in the Voice version than in either of the Agent or the Text version ($F(2,9) = 8.2$; $p = 0.009$). The mean values are indicated in Fig. 4. The result for the map area, while not statistically significant, shows a tendency toward a similar distribution of gaze points ($F(2,9) = 2.8$; $p = 0.11$). (For a comparison between gaze points in the agent and text box areas, see the Locked Attention Hypothesis below.)

Those results suggest that gaze points are not randomly distributed across the screen area but depend

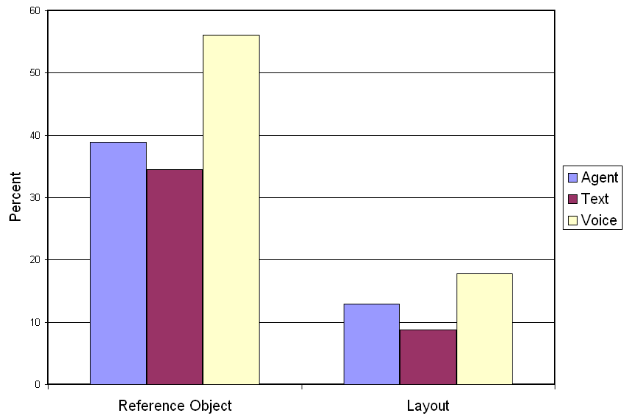


Figure 4: Impact of Agent vs. Text vs. Voice version on gaze points in reference object area and layout area.

on the presence or absence of a visible presentation medium. When an agent or a text box is present, users' attentional focus is more evenly shared between the presentation medium and the presented material.

**Locked Attention Hypothesis.** This hypothesis compares the portions that subjects focus on the agent (face or body) or the text box, which reveals text line by line. The mean for the agent is 18% of the total number of gaze points, and the mean for the text box is 32%. The $t$-test (one-tailed, assuming unequal variances) showed that subjects look significantly more often at the text box ($t(6) = -2.47$; $p = 0.03$).

This result can be seen as evidence that users spend considerable time for processing an object that gradually reveals new information. Locked attention can


Figure 5: “To your left is the layout of the apartment. As you can see, the apartment includes: bedroom, living room, dining room, den, kitchen and bathroom.”

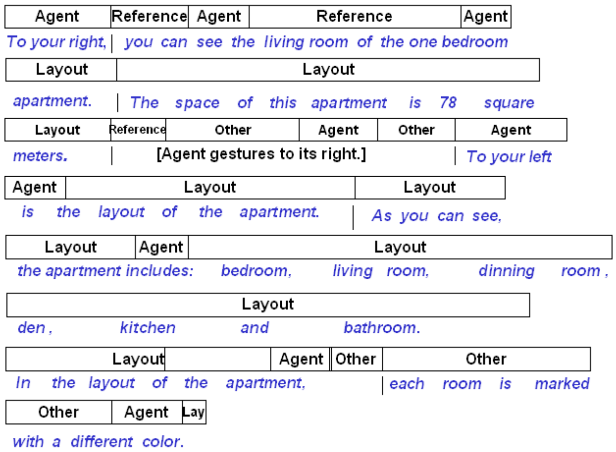


Figure 6: Example of attentional shifts in the introductory episode of the presentation.

prevent users from attending to other salient information (Faraday and Sutcliffe, 1996).

**Shift of Attention Hypothesis.** While a spatial analysis can indicate where attention is spent, it cannot reveal the nature of how users traverse the interface when watching a presentation. In order to address those more complex aspects of intelligent interfaces, we performed a (preliminary) *spatio-temporal* analysis of eye movement data. Figure 5 depicts a screen shot of the original view (taken by the outward directed camera of the EMR-8B system) of a subject in the Agent version. The dark colored dots are gaze points drawn by our program. The numbers have been added to the screen shot by hand. The frames around the agent (face, body) and the layout have been re-drawn for clarity. When the agent speaks the sentence in Fig. 5, the subject's focus of attention is first on the agent's face, next on the layout area, then it traverses back to the agent's face, and finally shifts to the layout area.

A more detailed description of one subject's attentional shifts is shown in Fig. 6. The rectangles above the sentences of the introductory episode of the apartment presentation indicate the focus of the subject's attention. The surface structure of the sentences is synchronized with attentional focus. Observe that the subject initially shifts attention between the agent and the living room (the reference object), and when the agent says “The space of this apartment is 78 square meters”, the subject focuses on the layout that depicts the size of the apartment. In the following, the subject partly attends to the agent's gesture, and after some occasional shifts to other areas, fixates on the layout. When the agent explains how the rooms are marked, the subject is apparently not attending to the layout during the utterance of the sentence.

The attentional shifts in the example of Fig. 6 suggest that users can perceive animated agents to possess a certain degree of competence, such as directing the user to locations of interest. Even more importantly, it demonstrates how a user re-directs attentional focus back to the agent after being directed to a reference object, which supports the interpretation of users expecting agents to provide them conversational cues and other meaningful information.

As a first attempt to provide a systematic spatio-temporal analysis of eye movements for intelligent embodied interfaces, we propose a Instructor–Reference–Instructor (IRI) triple as a basic unit for evaluation. An IRI denotes a situation where the user first attends to an instructor, a referring entity like an agent or a text box, then focuses on a reference object, and afterwards shifts attention back to the instructor. IRIs appear to be important interaction patterns in conversation, including direction-giving tasks (Nakano et al., 2003), and strong indicators of the instructor being conceived of as a social actor.

As a preliminary evaluation, we compared the number of IRIs of the Agent and Text versions for the episode displayed in Fig. 6 (plus one sentence). Here, both the living room and the layout qualify as reference objects. Figure 6, e.g., has 4 IRIs. The $t$-test on the small sample was not significant ($t(5) = 1.75$; $p = 0.07$). The means are: Agent (4.34) and Text (2). While this outcome indicates a tendency, further analysis with more episodes is needed to support the hypothesis that animated agents trigger conversational behavior in users.

**Agent Face–Body Hypothesis.** This hypothesis has been tested by summarizing gaze points that are contained in either the agent face or the agent body region. It could be shown that subjects were looking mostly at the agent's face (mean = 83.1%; stdev = 6.8), which supports the hypothesis that users interact socially with interface agents.

**Questionnaire.** The questionnaire contained two types of questions, one focusing on the subjects' general impression of the presentation, the other on the subjects' ability to recall shown items. In the first set of questions, subjects were asked (i) whether they would want to live in the apartment, (ii) whether they would recommend the apartment to a friend, and (iii) whether they thought the presentation helped them in their decision to rent the apartment. A 5 point Likert scale was used, ranging from "1" (strongly agree) to "5" (strongly disagree). The intention of questions (i) and (ii) was to investigate the effect of the presentation type on the users' perception of the apartment, but there were no results of statistical significance. An ANOVA of the third question, however, showed that subjects judged the Voice version to be more helpful that either of the other versions ($F$(2,12) = 8.9; $p$ = 0.004). The means are: Agent (2.2), Text (2.8), and Voice (1.2).

The second set of questions (eight in total) asked subjects for details of the presentation, such as "What could you see from the window in the living room?". Answers could be chosen from three options. The percentage of correct answers was 81.25% for the Agent version, 80% for the Text version, and 87.5% for the Voice version.

The results obtained from the questionnaire indicate that a presentation given by a disembodied voice is superior to an embodied agent or text together with underlying speech. This outcome supports the interpretation of agents carrying the risk of distracting users from the material being presented (van Mulken et al., 1998). On the other hand, agents might provide a more enjoyable experience to the user, but that dimension was not tested in the present study.

# 4 Conclusions

It is often argued that animated agents are endowed with *embodied intelligence* – they are able to employ human-like verbal and gestural behavior to behave naturally toward users (Cassell et al., 2000). However, little quantitative evidence exists that users also interact naturally with embodied agents in terms of involuntary indicators of interactivity such as attentional focus, which is an important prerequisite for their utility as virtual interaction partners. The same is true for the question to what extent users are involved in their interaction with embodied agents.

This paper has introduced a novel method for evaluating the interaction of users with animated interface agents, which is based on tracking users' eye movements. In terms of involvement in the interaction, this method allows us to evaluate whether users are *involved at the low level* and hence focus of the intended interface objects.

Primarily, it was demonstrated that the attentional focus hypothesized from gaze points constitutes a rich source of information about users' actual interaction behavior with computer interfaces. Both cumulative and temporal analyses of attentional focus revealed that users interact with animated agents in an essentially natural way. They follow the verbal and non-verbal navigational directives of the agent and mostly look at the agent's face. Unlike a textual interface (one revealing text line by line) that seems to capture users' attention to a high degree, users seem to attend to the visual appearance of the agent in a balanced way, with shifts to and from the object currently being presented. Although this result does not offer an interpretation as distinct as gaze behavior in grounding during face-to-face interaction (Nakano et al., 2003), it can provide valuable insights into the usability of the interface.

Besides an extended investigation of the obtained user gaze point data for spatio-temporal analysis, future work will also include the definition of comprehensive temporal measures of analysis for character-based interactive interfaces. A further interesting future direction is to track and analyze users' pupil dilating that has been shown as an index for confusion and surprise (Umemuro and Yamashita, 2003) and for affective interest (Hess, 1972).

# 5 Acknowledgements

We would like to thank Yukiko I. Nakano for her valuable suggestions on how to annotate the data, and Jin YingZi for helping with the analysis. This research was supported by the Research Grant (FY1999–FY2003) for the Future Program of the Japan Society for the Promotion of Science (JSPS).

# References

Elisbeth André, Jochen Müller, and Thomas Rist. The PPP Persona: A multipurpose animated presentation agent. In *Proceedings Advanced Visual Interfaces (AVI-96)*, pages 245–247. ACM Press, 1996.

Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors. *Embodied Conversational Agents*. The MIT Press, Cambridge, MA, 2000.

Roger M. Cooper. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, (6):84–107, 1974.

Doris M. Dehn and Susanne van Mulken. The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies*, (52):1–22, 2000.

Andrew T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, London, UK, 2003.

Pete Faraday and Alistair Sutcliffe. An empirical study of attending and comprehending multimedia presentations. In *Proceedings of ACM Multimedia 96*, pages 265–275, Boston MA, 1996.

Joseph H. Goldberg and Xerxes P. Kotval. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, 24:631–645, 1999.

Eckhard H. Hess. Pupillometrics: A method of studying mental, emotional and sensory processes. In N.S. Greenfield and R.A. Sternbach, editors, *Handbook of Psychophysiology*, pages 491–531. Holt, Rinehart & Winston, New York, 1972.

James C. Lester, Sharolyn A. Converse, Susan E. Kahler, S. Todd Barlow, Brian A. Stone, and Ravinder S. Bhogal. The Persona effect: Affective impact of animated pedagogical agents. In *Proceedings of CHI-97*, pages 359–366. ACM Press, 1997.

Tokyo Mansions, 2004. URL: `http://www.moveandstay.com`.

Microsoft. *Developing for Microsoft Agent*. Microsoft Press, Redmond, WA, 1998.

NAC. Image Technology, 2004. URL: `http://eyemark.jp`.

Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. Towards a model of face-to-face grounding. In *Proceedings of Association for Computational Linguistics (ACL-03)*, pages 553–561, 2003.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we know: Verbal reports on mental processes. *Psychological Review*, 84:231–259, 1977.

Toyoaki Nishida. Social intelligence design – an overview. In *New Frontiers in Artificial Intelligence. Joint JSAI 2001 Workshop Post-Proceedings*, pages 3–10. Springer, 2001.

Helmut Prendinger, Sylvain Descamps, and Mitsuru Ishizuka. MPML: A markup language for controlling the behavior of life-like characters. *Journal of Visual Languages and Computing*, 15(2):183–203, 2004.

Helmut Prendinger and Mitsuru Ishizuka. Introducing the cast for social computing: Life-like characters. In H. Prendinger and M. Ishizuka, editors, *Life-Like Characters. Tools, Affective Functions, and Applications*, Cognitive Technologies, pages 3–16. Springer Verlag, Berlin Heidelberg, 2004a.

Helmut Prendinger and Mitsuru Ishizuka, editors. *Life-Like Characters. Tools, Affective Functions, and Applications*. Cognitive Technologies. Springer Verlag, Berlin Heidelberg, 2004b.

Lei Qu, Ning Wang, and W. Lewis Johnson. Pedagogical agents that interact with learners. In *AAMAS-04 Workshop on Balanced Perception and Action in ECAs*, 2004.

Akikazu Takeuchi and Taketo Naito. Situated facial displays: Towards social interaction. In *Proceedings CHI 95 Conference*, pages 450–455, New York, 1995. ACM Press.

Hiroyuki Umemuro and Jun Yamashita. Detection of user's confusion and surprise based on pupil dilation. *The Japanese Journal of Ergonomics*, 39(4): 153–161, 2003.

Susanne van Mulken, Elisabeth André, and Jochen Müller. The Persona Effect: How substantial is it? In *Proceedings Human Computer Interaction (HCI-98)*, pages 53–66, Berlin, 1998. Springer.

Mark Witkowski, Yasmine Arafa, and Oscar de Bruijn. Evaluating user reaction to character agent mediated displays using eye-tracking technology. In *Proceedings AISB-01 Symposium on Information Agents for Electronic Commerce*, pages 79–87, 2001.