Understanding the Effect of Life-Like Interface Agents Through Users' Eye Movements

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

Chunling Ma, Jin Yingzi,
Arturo Nakasone, Mitsuru Ishizuka
Dept. of Information and Communication Eng.
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
ishizuka@miv.t.u-tokyo.ac.jp

ABSTRACT

We motivate an approach to evaluating the utility of life-like 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 a user's interaction experience on a moment-by-moment basis in addition to a cumulative (spatial) analysis of the user's areas of interest. We describe an empirical study in which we compare attending behavior of subjects watching the presentation of an apartment by three types of media: an animated agent, a text box, and speech only. The investigation of users' eye movements reveals that agent behavior may trigger natural and social interaction behavior of human users.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Human Factors

Keywords

User study, eye tracking, animated interface agents, web-based presentation

1. INTRODUCTION

Life-like 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. [1] 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 [4, 21]. While significant progress has been made in individual aspects of the 'life-likeness' 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 [13, 28]. Others investigated the likeability of different types of synthetic interface agents [14].

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 [6] pointed out, the broad variety of realizations of life-like 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 [18].

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 [7]. 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 life-like 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 tracked and analyzed eye movements while users were following the web page based presentation of different rooms

ICMI'05, October 4–6, 2005, Trento, Italy.

Figure 1: Life-like animated interface agent (left) and text box (right).

of an apartment. Three types of presentations were contrasted (see Fig. 1):

1. A life-like interface agent presents the apartment using speech and gestures;
2. The apartment is presented by means of a text box and read out by speech; and
3. 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 and as an input modality. The core part of the paper (Sect. 3) is devoted to the description of an experiment that provides both spatial and (preliminary) temporal analyses of users' eye movements during a presentation. Section 4 discusses the results of the study and Sect. 5 concludes the paper.

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 [8, 9, 24]. In the interactive use, a system responds to the observed eye movements and can thus be seen as an input modality [12, 7, 17].

An analysis of eye movements in order to assess the usability of an interface for a simple drawing tool was performed in [9]. 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. A study that analyzes the duration of eye fixations to determine the usability of different graph designs can be found in [24].

Attentional processing and comprehension of multimedia presentations is investigated by [8]. Core findings of the authors relevant to our domain (that will be partly tested in the study reported in Sect. 3) 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 [5] 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 [7, p. 167]).

The work of [30] employs 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 [15] 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, [30] observed that interface agents do attract the attention of users. Similar results have been obtained in [25] that compares an interface featuring either a (facial) agent or an arrow.

The study in [11] examines the effect of an animated agent and different voice types on comprehension and attention performance. While the agent was able to direct users' attention and maintain their engagement, no increased learning of the multimedia presentation could be demonstrated.

Besides their diagnostic role, eye movement data have also been used as an additional input modality to human–computer interaction. [12] investigates eye-based interaction techniques such as (interface) object selection, moving of an object (a variation of the 'drag-and-drop' operation) and scrolling of text. In the realm of life-like agent based systems, [23] consider a user's focus of attention (among others) to decide an appropriate response for an educational software, and [17] investigate attentional focus (among others) for a direction-giving task.

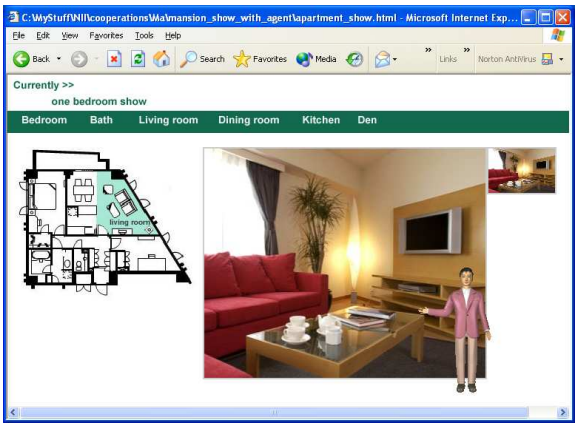

Figure 2: A life-like animated agent presents the living room.

3. METHOD

3.1 Experimental Design

A presentation of an apartment located in Tokyo has been prepared using a web page based interface [26]. 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 implemented for the experiment:

1. *Agent (& speech) version.* A character called "Kosaku" presents the apartment using synthetic speech and deictic facial and hand gestures (see Fig. 2). Only simple "left"/"right" gestures (rather than full 360 degrees pointing) is available to the character, which is controlled by a version of MPML [20].
2. *Text (& speech) version.* The presentation content of each scene is displayed by a text box and read out by Microsoft Reader (see Fig. 1, right).
3. *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.

It is important to mention that the presentation interface does not involve active interaction. However, we argue that users watching a presentation *interact* – even involuntarily – by their eye movement activity. Evidence for this claim will be provided below.

3.2 Subjects

Fifteen subjects (3 female, 12 male), all students or staff from the Univ. of Tokyo, participated in the study, with five subjects randomly assigned to each version. (Similar to other eye tracking experiments, the rather small number of subjects was necessitated by the expensive data analysis.) The age of subjects 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 (the EMR monitor) was used to control the eye tracking system, a NAC Image Technology Eyemark Recorder [16]. The eye mark recorder is shown in Fig. 3 and the experimental setup is shown in Fig. 4.

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 and allowed to synchronize eye-tracking recording and video recording.

Figure 3: NAC EMR-8B.

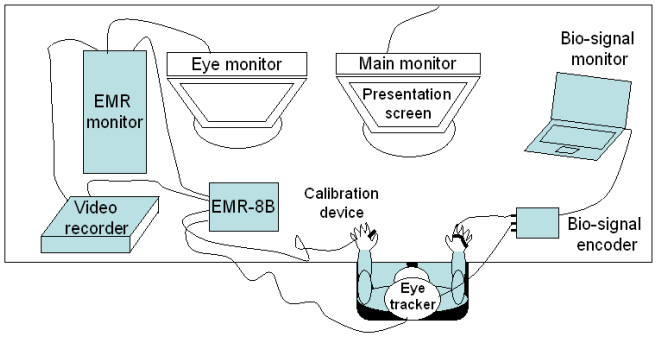

Figure 4: Experimental setup.

When eye movements are relatively steady for a short period (250–300 ms), they are called *fixations* whereas rapid shifts from one area to another are called *saccades* [12]. During a saccade, no visual processing takes place. If a cluster of gaze points has less than 6 entries, it is categorized as part of a saccade [9], i.e. we assume a minimum duration of 100 ms for a fixation (at 60 Hz). In the present study, fixations and saccades are defined wrt. screen areas only.

The subjects were also connected to a bio-signal encoder that provides skin conductance and heart rate sensors. The bio-signal part of the experiment did not yield significant results in terms of the arousal level of subjects or the valence of their emotional perception, and 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. E.g. in Fig. 2, 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:

1. 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).
2. The area of the reference object is the smallest rectangle demarcating the object currently described.
3. The map or layout area (a designated, permanent reference object) is the field on the screen that displays the layout (map) of the room.
4. Other screen areas.

Our program first maps eye data to xy-coordinates of the video sequence and then counts the gaze points in each of the four categories. All data accounted for in the analysis are derived from the activity of subjects' left eyes. In each version, data of one subject had to be discarded due to technical problems.

3.6 Results of Spatial Analysis

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* (or cumulative) analysis counts the gaze points that fall within certain screen areas and hypothesizes areas of interest. Spatio-temporal analysis will be discussed below.

3.6.1 Focus of Attention Hypothesis

In order to support the Focus of Attention Hypothesis, we specifically investigate the reference object area and the layout (map) 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

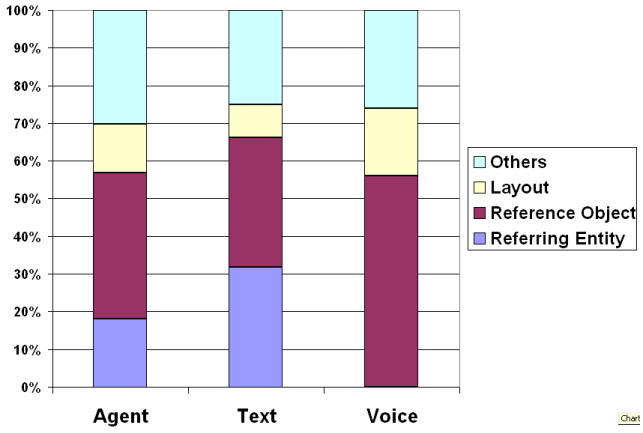

Figure 5: Impact of Agent vs. Text vs. Voice version on gaze points in different screen areas.

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 level of statistical significance is set to 5%.) The percentual proportions are indicated in Fig. 5. 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.)

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

3.6.2 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% (see Fig. 5). 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 prevent users from attending to other salient information [8].

3.6.3 Agent Face–Body Hypothesis

The Agent Face–Body 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 can be interpreted as supportive evidence for the hypothesis that users interact socially with life-like interface agents [30].

This result begs the question whether subjects were aware of the deictic arm gestures, which is obviously essential to their effectiveness. Since data were not analyzed at this granularity level, we can only report on our (non-systematic)

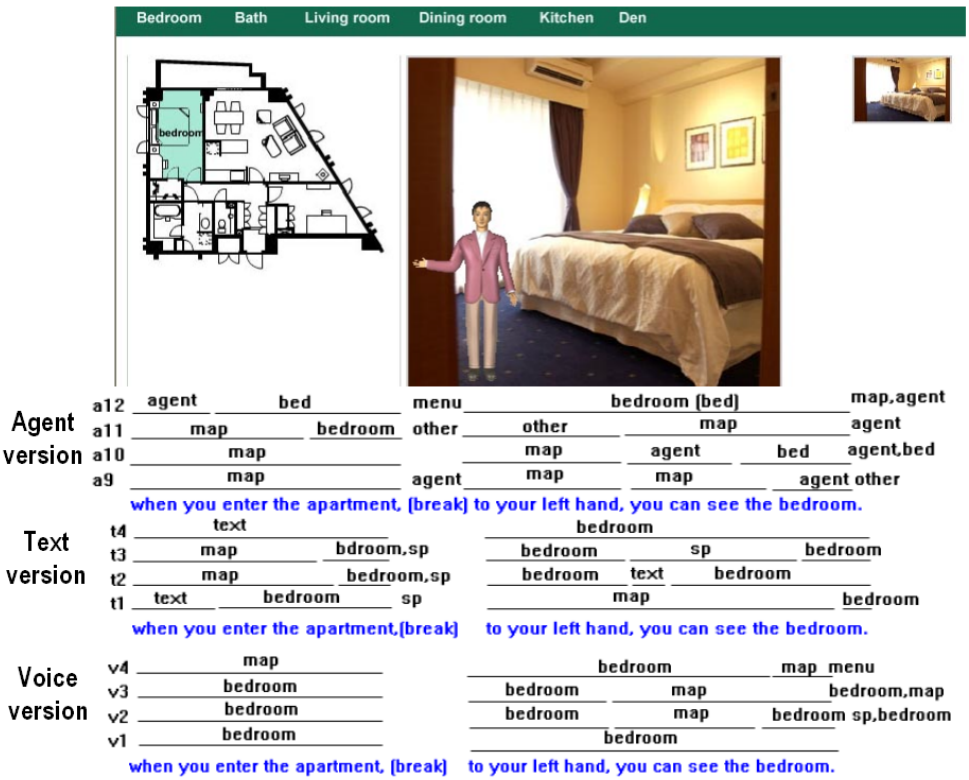

Figure 6: Effect of deictic reference on eye movement. Each row of underlined text shows the gaze locations of subjects denoted by a9, a10, . . . , t1, t2, . . . ("sp" refers to the small picture to the top-right.)

observations while looking at the videos. When the agent performs an arm gesture, subjects' attention is attracted by the animation change for a very short time and their gaze subsequently often 'slides' along the agent's arm in the direction of the reference object.

3.7 Results of Spatio-Temporal Analysis

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 multi-modal and multimedia interfaces, we performed a (preliminary) *spatio-temporal* analysis of eye movement data with twenty-two sentences. In the following, we present our observations.

3.7.1 Auditory Language Processing

We first discuss the Auditory Language Processing Hypothesis with respect to our three conditions. In Fig. 6, the referring entity (agent, text box, voice) is intended to direct the user's attention to the map (layout) area that depicts the bedroom. It is important to notice that unlike the study in [5], the word "bedroom" in the uttered sentence is not unambiguous with regard to its reference object: "bedroom" might refer to either the specified area in the map (layout) to the left or to the picture of the bedroom to the right. In the Agent version only, subjects mostly direct their attention to the intended direction, the map. Although subjects in the Voice version eventually attend to the map, subjects in the Agent version (mostly) do so from the beginning. This kind of user behavior is seemingly affected by the agent performing an according deictic gesture (to its right) shortly before starting the utterance. In the Text version, subjects seemingly cannot resolve the reference since most subjects focus on the unintended reference object (the picture of the bedroom). A similar eye movement pattern was observed in comparable other utterances.

The sentence in Fig. 7 is similar to the sentences used in [5] as it contains a 'trigger word' – here the word "window" that is both spoken and has a semantically related visualization (the picture of a window). In the Agent version subjects focus on the visual window while or shortly after they hear the word "window". (One subject already looks at the window before it is uttered.) A likely reason is that the agent performs a deictic (facial) gesture in that it turns its head to the relevant direction. The Voice version does not show a clear focus pattern of subjects' eye movements. In line with the Locked Attention Hypothesis, subjects in the Text version first read the whole sentence in the text box, and then direct their attention to the picture of the window.

3.7.2 Instructor–Reference–Instructor Triples

As a first attempt to provide a systematic spatio-temporal analysis of eye movements for interfaces with navigational aids, we propose an 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 [17], and indicators of the instructor being conceived of as a social actor.

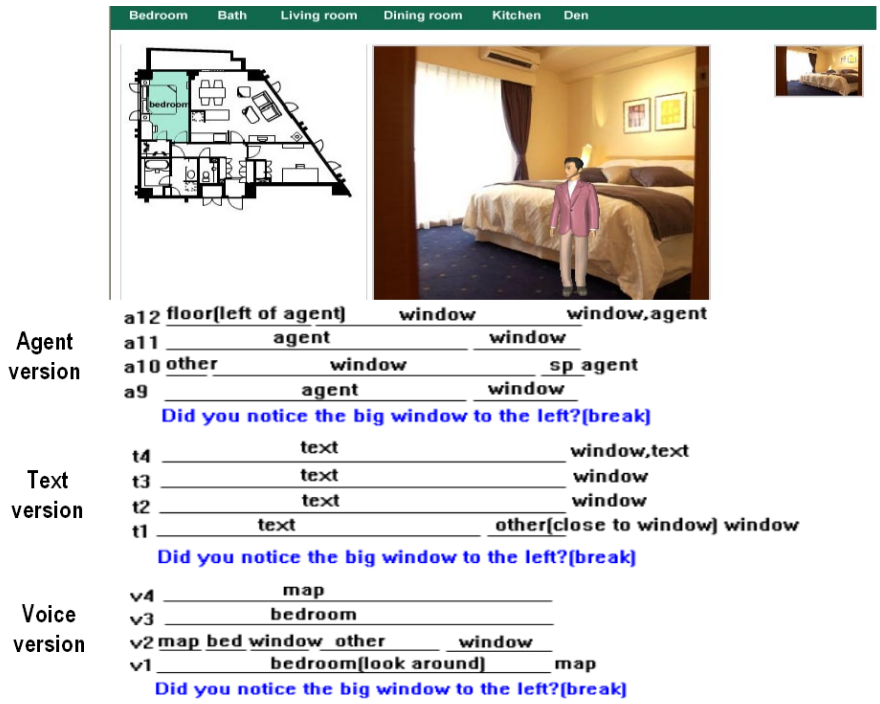

Figure 7: Effect of auditory language processing and deictic reference on eye movement.

Table 1: Shift of attentional focus at sentence breaks and referential acts.

Agent version	**a9**	**a10**	**a11**	**a12**
To agent at sentence break	50%	54%	45%	40%
To reference object	75%	85%	73%	58%
Text version	**t1**	**t2**	**t3**	**t4**
To text at sentence break	32%	50%	18%	27%
To reference object	50%	64%	55%	72%

A representative example is the situation where the agent utters: "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." Here, subjects often initially shift 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", subjects first focus on the layout (map) that depicts the size of the apartment, subsequently partly attend to the agent's gesture, and eventually fixate on the layout.

The attentional shifts suggest that subjects 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 redirects 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. This hypothesis is also supported by the fact that users sometimes focus on the agent during breaks between sentences or sentence parts, seemingly waiting for the agent (that holds the floor) to continue.

Table 1 (upper part, Agent version) shows the percentages that subjects (a9, ..., a12) redirect their attentional focus (back) to the agent after sentence breaks, and those where subjects could precisely shift to the reference object referred to by the agent. The percentages for the Text version are given in the lower part of Table 1. The t-test indicates that in the Agent version, subjects look back to the instructor at sentence breaks significantly more than in the Text version ($t(6) = 2.09; p = 0.05$), with a tendency for more accurately shifting their attention to the reference object ($t(6) = 1.67; p = 0.07$).

3.7.3 *Summary of Observations*

Here we briefly summarize our initial findings (based on twenty-two sentences) about the three types of media:

- An agent's referential (arm or facial) gestures may direct the user's focus of attention to the intended reference object better than a text box or only voice.
- If the uttered sentence contains a trigger word – a word that has a corresponding semantically related visualization – an agent using gestures helps users to locate the (visual) reference object quickly. By contrast, directional support by a text box or voice often shows considerable latency.
- Users often redirect their attention back and forth between the animated agent and the reference object, similar to human–human communication.

3.8 Questionnaire Results

In addition to physiological user data, we also analyzed questionnaires as a standard interface evaluation method. 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:

1. Whether they would want to live in the apartment;
2. Whether they would recommend the apartment to a friend; and
3. 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 (1) and (2) 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 can be superior to an agent or text together with underlying speech in terms of perceived helpfulness.

4. DISCUSSION

This paper has introduced a novel method for evaluating the interaction with life-like interface agents, which is based on tracking users' eye movements, an objective method that does not distract the user from the primary task. Although eye tracking has been abundantly used in psychology, multimedia, and related studies [7], its application to human–agent interaction is currently rare.

The study has 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 life-like interface agents in an essentially natural way. Users 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 captures 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. This observation also forwards the discussion about the believability of life-like agents in a new way. The eye movements of users watching a presentation given by an agent provide quantifiable evidence of their perception of the agent's believability. Here, the believability of the agent can be conceived as its ability to effectively direct the user's focus of attention to objects of interest.

A sometimes heard concern about employing eye tracking technology to evaluate the effect and utility of animated interface agents is that most of the results were to be expected. With the exception of the related study described in [30], our work is the first that aims at investigating the effect of animated agent behavior on a moment-to-moment basis. The aforementioned expectation is seemingly based on the assumption that even on the mostly involuntary level of eye movements, humans would interact with an animated presenter as they do with a real human presenter. This assumption, in our view, is considerably stronger than assuming the often reported "suspense of disbelief" when interacting with virtual figures [2], and hence, worth investigating.

A natural extension of our work is to explore eye movements in the context of human–agent interaction where the user may actively participate in the conversational process. [17] designed a life-like agent (Mack) that provides the user with directions on a (shared) physical map, and derives information about the user's conversational state from gaze behavior. For instance, if the user is gazing at the shared referent (the map), it is interpreted as positive evidence of understanding on the part of the user, i.e. the information is assumed as 'grounded'.

Besides eye movement data, we also collected biometric user information in order to study the affective state of user during the presentation. However, contrary to the study described in [29], neither skin conductance nor heart rate activity yielded significant differences between the presentation conditions. The outcome of the questionnaire supports the interpretation of life-like agents carrying the risk of distracting users from the material being presented (see also [28]). On the other hand, agents might provide a more enjoyable experience to the user, but that dimension was not tested in the present study.

5. CONCLUSIONS

It is often argued that life-like agents are endowed with *embodied intelligence* – they are able to employ human-like verbal and gestural behavior to behave naturally toward users [4]. However, so far little quantitative evidence exists that users also interact naturally with animated agents in terms of largely involuntary characteristics of interactivity such as attentional focus, which is an important prerequisite for their believability and utility as virtual interaction partners. The study presented in this paper demonstrated that life-like agents may trigger natural behavior in users.

Besides an extended investigation of the microstructure of gaze transitions, future work will also include the definition of comprehensive temporal measures of analysis for agent based interactive interfaces. Here, the work described in [9] may serve as a starting point. A further interesting future direction is to track and analyze users' pupil dilation that has been shown as an index for confusion and surprise [27] and for affective interest [10, 19].

In terms of the future of interfaces employing life-like agents, the study in this paper is intended to motivate and propel research into agent based interfaces that recognize physiological information of users in real-time, and respond appropriately to users' affective state and attentional focus (see [22] for an early attempt). It is our hope that complementing multi-modal output and synchronization of behavior of life-like agents by multi-sensor input recognition and signal fusion [3] will greatly advance interfaces that realize effective, efficient, and natural communication between humans and computers.

Acknowledgements

We would like to thank Yukiko I. Nakano for her valuable suggestions on how to annotate the data. This research was

supported by the Research Grant (FY1999–FY2003) for the Future Program of the Japan Society for the Promotion of Science (JSPS) and by a JSPS Encouragement of Young Scientists Grant (FY2005–FY2007).

6. REFERENCES

[1] E. André, J. Müller, and T. Rist. The PPP Persona: A multipurpose animated presentation agent. In *Proceedings Advanced Visual Interfaces (AVI-96)*, pages 245–247. ACM Press, 1996.

[2] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

[3] W. Burleson, R. Picard, K. Perlin, and J. Lippincott. A platform for affective agent research. In *Proceedings 3rd International Conference on Autonomous Agents & Multi Agent Systems (AAMAS-03)*, New York, 2004. ACM Press.

[4] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. The MIT Press, Cambridge, MA, 2000.

[5] R. 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.

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

[7] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, London, UK, 2003.

[8] P. Faraday and A. Sutcliffe. An empirical study of attending and comprehending multimedia presentations. In *Proceedings of ACM Multimedia 96*, pages 265–275, Boston MA, 1996.

[9] J. H. Goldberg and X. P. Kotval. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, 24:631–645, 1999.

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

[11] C. Hongpaisanwiwat and M. Lewis. Attention effect of animated character. In *Proceedings Human–Computer Interaction (INTERACT-03)*, pages 423–430. IOS Press, 2003.

[12] R. J. K. Jacob. The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems*, 9(3):152–169, 1991.

[13] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal. The Persona effect: Affective impact of animated pedagogical agents. In *Proceedings of CHI-97*, pages 359–366. ACM Press, 1997.

[14] H. McBreen, P. Shade, M. Jack, and P. Wyard. Experimental assessment of the effectiveness of synthetic personae for multi-modal e-retail applications. In *Proceedings 4th International Conference on Autonomous Agents (Agents'2000)*, pages 39–45, New York, 2000. ACM Press.

[15] Microsoft. *Developing for Microsoft Agent*. Microsoft Press, Redmond, WA, 1998.

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

[17] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of Association for Computational Linguistics (ACL-03)*, pages 553–561, 2003.

[18] R. E. Nisbett and T. D. Wilson. Telling more than we know: Verbal reports on mental processes. *Psychological Review*, 84:231–259, 1977.

[19] T. Partala and V. Surakka. Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, 59:185–198, 2003.

[20] H. Prendinger, S. Descamps, and M. Ishizuka. MPML: A markup language for controlling the behavior of life-like characters. *Journal of Visual Languages and Computing*, 15(2):183–203, 2004.

[21] H. Prendinger and M. Ishizuka, editors. *Life-Like Characters. Tools, Affective Functions, and Applications*. Cognitive Technologies. Springer Verlag, Berlin Heidelberg, 2004.

[22] H. Prendinger and M. Ishizuka. The Empathic Companion: A character-based interface that addresses users' affective states. *International Journal of Applied Artificial Intelligence*, 19(3):267–285, 2005.

[23] L. Qu, N. Wang, and W. L. Johnson. Pedagogical agents that interact with learners. In *AAMAS-04 Workshop on Balanced Perception and Action in ECAs*, 2004.

[24] J. Renshaw, J. Finlay, D. Tyfa, and R. Ward. Understanding visual influence in graph design through temporal and spatial eye movement characterisitics. *Interacting with Computers*, 16:557–578, 2004.

[25] A. Takeuchi and T. Naito. Situated facial displays: Towards social interaction. In *Proceedings CHI 95 Conference*, pages 450–455, New York, 1995. ACM Press.

[26] Tokyo Mansions, 2004. URL: `http://www.themansions.jp/`.

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

[28] S. van Mulken, E. André, and J. Müller. The Persona Effect: How substantial is it? In *Proceedings Human Computer Interaction (HCI-98)*, pages 53–66, Berlin, 1998. Springer.

[29] G. Wilson and M. Sasse. Listen to your heart rate: Counting the cost of media quality. In A. Paiva, editor, *Affective Interactions – Towards a New Generation of Computer Interfaces*, pages 9–20. Springer, Berlin Heidelberg, 2000.

[30] M. Witkowski, Y. Arafa, and O. 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.