Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Interacting with Computers 19 (2007) 281–292

Interacting with Computers

www.elsevier.com/locate/intcom

Eye movements as indices for the utility of life-like interface agents: A pilot study

Helmut Prendinger [a,*], Chunling Ma [b], Mitsuru Ishizuka [b]

[a] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*

[b] *Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*

Received 9 August 2005; received in revised form 22 July 2006; accepted 15 October 2006
Available online 28 November 2006

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 without distracting from the primary task. 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 a pilot study in which we compare attending behavior of subjects watching the presentation of a computer-generated apartment layout and visualization augmented by three types of media: an animated agent, a text box, and speech only. The investigation of eye movements revealed that deictic gestures performed by the agent are more effective in directing the attentional focus of subjects to relevant interface objects than the media used in the two control conditions, at a slight cost of distracting the user from visual inspection of the object of reference. The results also demonstrate that the presence of an interface agent seemingly triggers natural and social interaction protocols of human users.

Keywords: Animated interface agents; Evaluation; Eye movements; Focus/shift of attention; 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. 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, 2004). 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 (Lester et al., 1997; van Mulken et al., 1998) but not on performance (Craig et al., 2002). Others investigated the likeability of different types of synthetic interface agents (McBreen et al., 2000). Moreno (2004) contrasts the beneficial and disadvantageous effects of animated agents within the context of multimedia presentations for e-learning.

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, engagement, utility, and ability to attract attention. However, as Dehn and van Mulken (2000) pointed out, the broad variety of realizations of life-like agents and

[*] Corresponding author. Tel.: +81 3 4212 2650; fax: +81 3 3556 1916.
E-mail addresses: helmut@nii.ac.jp (H. Prendinger), ishizuka@i.u-tokyo.ac.jp (M. Ishizuka).

0953-5438/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2006.10.004

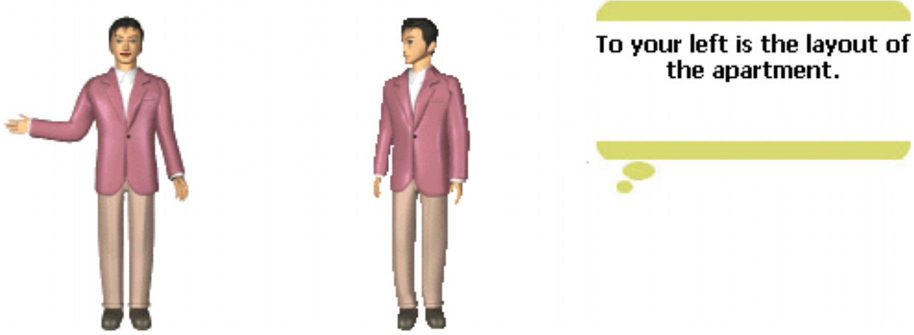

Fig. 1. Life-like animated interface agent performing deictic arm–hand gesture (left), deictic facial gesture (middle), and text box (right).

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).

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.[1] 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). Specifically, 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 to intended interface objects?

Hence, eye movement data might 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.

In our study, we tracked and analyzed eye movements while users were following the Web page based presentation of different rooms of an apartment. Three types of presentations were contrasted:

1. A life-like interface agent presents the apartment using speech and deictic arm–hand gestures (Fig. 1, left) or deictic facial gestures (Fig. 1, middle);
2. The apartment is presented by means of a text box and read out by speech (Fig. 1, right); and
3. The presentation is commented by speech only.

It is important to mention that the presentation interface does not involve active interaction in the sense that users would be able to control the interface in some way. However, we argue that users merely watching a presentation *interact* – even involuntarily – by their eye movement activity. We will provide ample evidence for this claim.

The remainder of the paper is organized as follows. The next section overviews work related to using eye movement and other physiological signals as an evaluation method for user interfaces and as an input modality. The core part of the paper (Section 3) is devoted to the description of a pilot study that provides both spatial and temporal analyses of subjects' eye movements during a presentation. Section 4 discusses the results of the study and Section 5 concludes the paper.

2. Related work

This section reports on work that employs eye movements or bio-signals in the context of user interfaces. Eye movement data have been analyzed for two main purposes, *diagnostic* and *interactive*. In 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; Renshaw et al., 2004). In interactive use, a system responds to the observed eye movements and can thus be seen as an input modality (Jacob, 1991; Duchowski, 2003; Nakano et al., 2003). The first part of this section mainly focuses on the diagnostic use of eye movements. A similar distinction can be drawn for the case of bio-signals (Picard, 1997). In the second part of this section we will again put emphasis on the diagnostic use of those signals.

2.1. Attention tracking

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 was efficient scanning behavior, i.e. a short scan path to the target object. The chief merit

[1] This manuscript is a significantly improved and extended version of Prendinger et al. (2005a).

of this study was 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 the investigation of the effect of deictic references of animated agents to interface objects. A related study analyzing the duration of eye fixations to determine the usability of different graph designs can be found in Renshaw et al. (2004), which demonstrated the importance of the location of the legend of a graph and its spatial relationship to the area where data are displayed.

While these two studies were concerned with a static interface configuration, we now turn to discussing studies that are based on dynamic content. Faraday and Sutcliffe (1996) investigated attentional processing and comprehension of dynamically changing multimedia presentations. Core findings of the authors relevant to our domain (that will be partly tested in the study reported in Section 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 Cooper (1974) who (successfully) tested the hypothesis: "When people are simultaneously presented with spoken language and a visual field containing elements semantically related to the informative items of speech, they tend to spontaneously direct their line of sight to those elements which are most closely related to the meaning of the language currently heard" (Cooper, 1974, p. 85).

An alternative way to test a user interface based on eye movements has been proposed in Lin et al. (2004). In this study, a so-called "hand-eye" measure is advocated that evaluates the interface not only by eye behavior reflecting a user's cognitive processes, but also by their outcome, a particular physical action of the user (a mouse click).

The work most closely related to ours has been performed by Witkowski et al. (2001) who 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 conjectured that the agent will direct the attention of the users to the item of interest (help buttons and pictures of wines), following the agent's verbal comments. However, the results of their study did 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 showed that users mostly focused on reading the text, rather than attending to the agent or to the depicted 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, the face was focussed on the most. In general, Witkowski et al. (2001) observed that interface agents did 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 a graphical arrow.

Hongpaisanwiwat and Lewis (2003) investigated the effect of an animated agent and different voice types on comprehension and attention. The authors showed that the agent was able to direct users' attention and maintain their engagement, but no increased learning of the multimedia presentation could be demonstrated. Although the study examined the attentional effect of animated agents, eye movement data were not used as an evaluation methodology.

Besides its diagnostic role, eye movement information has also been used as an additional input modality, in order to increase the bandwidth in human–computer interaction. Jacob (1991) investigated 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 Oyekoya and Stentiford (2004), eye information is used for predicting users' interest in an image retrieval task. In the realm of life-like agent based systems, Qu et al. (2004) considered a user's focus of attention (among others) to decide an appropriate response for an educational software, and Nakano et al. (2003) investigated attentional focus (among others) for a direction-giving task.

2.2. Emotion and stress tracking

Complementary to studies targeted at the analysis of user attention, which is related to the cognitive aspect of interacting with computers, a growing body of experimental research targets usability from the viewpoint of user emotions that are elicited during interaction. These studies are covered under the umbrella term of 'affective computing' (Picard, 1997), a research area concerned with recognizing emotion and developing stress-reducing strategies for negatively aroused interface users (Klein et al., 2002). While attention and emotion (or more generally, affect) are clearly distinct concepts, both relate to mostly non-conscious experiences of a computer user that might affect the usability and utility of the interface. With regard to our study, we were interested in testing the wide-spread assumption that animated agents provide a more natural medium of information delivery (Rist et al., 2004), which might be demonstrated by a stress-reducing effect due to their presence.

In the literature on emotion (affect), there is ample evidence that physiological signals (or bio-signals) such as skin conductance, muscle tension, and heart rate provide important information regarding the intensity and quality of a person's experience, and can thus be used to infer a user's affective state (arousal and valence of feeling) or even emotion (Picard, 1997; Levenson, 2003). Since most studies investigate either attention or emotion, we discussed eye movements separately from other physiological signals, e.g. skin conductance or heart rate. The latter type of physiological response will also be called biometric signals or bio-signals.

Wilson and Sasse (2000) investigated the use of biometric data in order to assess user cost of the reception of different levels of multimedia quality in the context of Internet-based video-conferencing. User cost is operationalized by the following bio-signals indicating stress: blood volume pulse, heart rate, and galvanic skin resistance. In the study, participants were presented interviews at alternating frame rates, 5–25–5 fps and 25–5–25 fps. The authors could demonstrate that 75% of the subjects showed significantly increased stress levels at 5 frames per second, while only 16% of the subjects noticed that the frame rate had changed. Hence, it could be shown that physiological and subjective responses to multimedia do not always correlate.

A comparable study has been conducted by Ward and Marsden (2003) who asked their subjects to perform a task with two types of websites, one well-organized and one poorly organized, with distracting features such as many pull-down menus and pop-up windows. Although results were not statistically significant, subjects interacting with the poor Web design showed increased levels of skin conductance, heart rate, and finger blood volume.

3. Methods

3.1. Experimental design

A presentation of an apartment located in Tokyo has been prepared using a Web page based interface Tokyo 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. Fig. 2 depicts a situation where the agent refers to the living room. Besides the pictures located in the center, the interface shows the layout (or map) of the apartment to the left, a miniature version of the center picture to the right, and other interface objects at the top. For simplicity, the last two mentioned interface elements were not considered in the analysis.

Three versions of the apartment show have been implemented for the pilot study:

Agent (& Speech) version: A character called "Kosaku" presents the apartment using synthetic speech and deictic gestures (see Fig. 2). Only simple "left"/"right" arm–

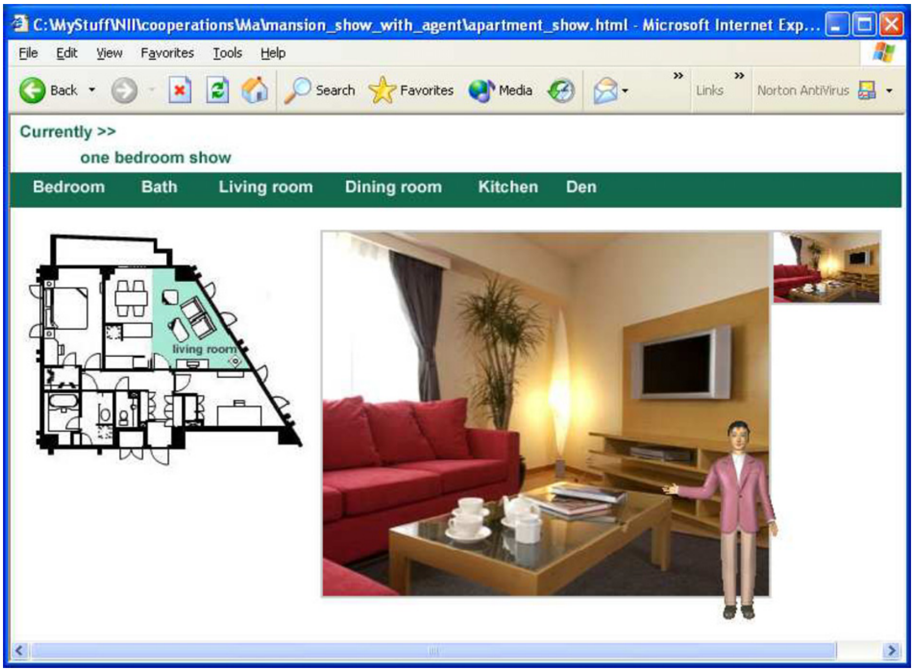
C:\MyStuff\NII\cooperations\Ma\mansion_show_with_agent\apartment_show.html - Microsoft Internet Exp...
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Links
Norton AntiVirus
Currently >>
one bedroom show
Bedroom
Bath
Living room
Dining room
Kitchen
Den
living room

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

hand and facial gestures rather than full 360° pointing are available to the character (see Fig. 1, left and middle, respectively). The animation of the character was developed by Hottolink (2002) and is controlled by a version of the Multi-modal Presentation Markup Language (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 (see Fig. 1, right).

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.[2] The same type and speed of (synthetic) voice was used in all versions.

3.2. Hypotheses

Based on previously discussed related research, we formulate the following hypotheses covering various aspects of the utility of life-like interface agents.

Hypothesis 1. (H1) *Focus of Attention Hypothesis*: We predict that gaze points are not randomly distributed across the interface, but depend on the presentation condition. Since the Agent and Text versions provide a visual medium for narration (an agent, a text box), less attention will be spent on the referred screen objects than in the Voice version.

While this hypothesis might sound trivial, there are currently no results concerning the proportions that users attend to the presenting entity (if visible) and the other interface areas.

Hypothesis 2. (H2) *Locked Attention Hypothesis* (Faraday and Sutcliffe, 1996; Witkowski et al., 2001): Since the presentation comments in the Text version are revealed line by line (and hence changing), the attention of subjects will be 'locked' by reading the text.

Hypothesis 3. (H3) *Agent Face–Body Hypothesis* (Witkowski et al., 2001): Similar to human–human communication, subjects spend significantly more time with attending to the face of the agent, than to its body.

Hypothesis 4. (H4) *Extended Auditory Language Processing Hypothesis*: The prediction is that, compared to the Text and Voice versions, the deictic gestures of the agent can more effectively direct the attention of the user to the referenced area of the interface.

"Extended" here means that the hypothesis originally formulated for spoken language and corresponding visual objects (Cooper, 1974) is extended to the presence of an animated agent's deictic gestures.

Hypothesis 5. (H5) *Social Interaction Protocol Hypothesis* (Nakano et al., 2003): Subjects in the Agent version shift their focus of attention back to the agent after they have been directed to a particular interface object.

This hypothesis can be seen as a 'dynamic' version of Hypothesis (H3).

Hypothesis 6. (H6) *Cost of Media Hypothesis* (Rist et al., 2004): Since animated agents are a natural medium of information presentation, subjects will be less stressed in the Agent version than in the other two conditions.

Hypotheses 1–3 are tested by a spatial (cumulative) analysis of eye movement data, whereas Hypotheses 4 and 5 are based on a spatio-temporal (dynamic) analysis. Hypothesis 6 will be tested by analyzing biometric data (skin conductance and heart rate). Other hypotheses regarding participants' subjective perception of the different types of media and recall of presented information will be based on the analysis of a questionnaire.

3.3. Subjects

Fifteen subjects (3 females and 12 males), all students or staff from the University of Tokyo, participated in the study, whereby five subjects were 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 1000 Yen for participation. In three 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.

3.4. Apparatus

The presentation of the apartment was hosted on a computer with a 17 in. (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 (NAC, 2004). 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 subjects' 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 subjects' visual field, including the Main monitor. The system has a sampling rate of 60 Hz. The head posture of subjects was maintained with a chin rest, with the eyes at a distance of 24 in. (60 cm) from the Main monitor. A digital video recorder capturing the data from the third camera was

[2] A condition for voice and pointing by an arrow could have been included, but was not implemented in the current study.

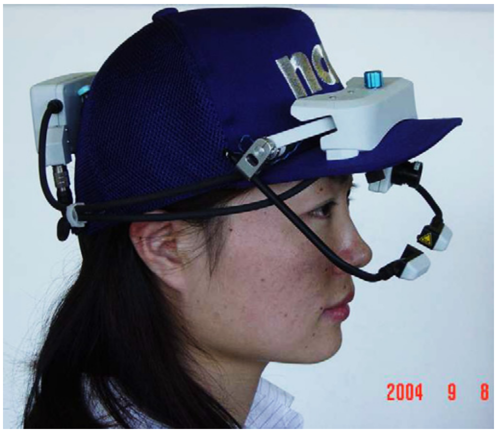

Fig. 3. NAC EMR-8B.

connected to the computer that processed the eye movements and allowed us to synchronize eye-tracking recording and video recording.

The subjects were also connected to a bio-signal encoder (designed by Hiroshi Dohi from the University of Tokyo) that provides skin conductance (SC) and heart rate (HR) sensors. The sampling rate was set to 2 samples per second. The SC sensing system is built into an elastic band and put over the subject's palm of their right hand. The HR sensor is part of an ear-clip and is attached to the subject's right ear. The encoder was connected to a third computer. Signal data were exported to an Excel spreadsheet.

3.5. 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 then attached to the bio-sensors and subsequently put on the cap with the eye tracker. Calibration was performed by instructing the subject to fixate nine points in the screen area. After calibration was completed, subjects were asked to relax for a period of 3 min that served as the baseline period for skin conductance and heart rate values. After that, the subjects were shown the presentation that lasted for 8 min.

After the presentation, subjects were freed from the eye and bio-signal 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.

3.6. Data analysis

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* (Salvucci and Goldberg, 2000). During a saccade, no visual processing takes place. If a cluster of gaze points has less than six entries, it is categorized as part of a saccade (Goldberg and Kotval, 1999). In the present study, no fixation (or saccade) algorithm was required since we were only interested in gaze points within the predefined areas described below (that were above the mentioned threshold of six entries), and the temporal evolution of a user's attention to those areas.

We wrote a computer program that first maps gaze points to *xy*-coordinates of the video sequence and then counts the number of 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, eye data of one subject had to be discarded due to technical problems.

For analysis, the recorded video data of each 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. For example, in Fig. 2, the scene consists of the agent performing a arm–hand gesture to its right

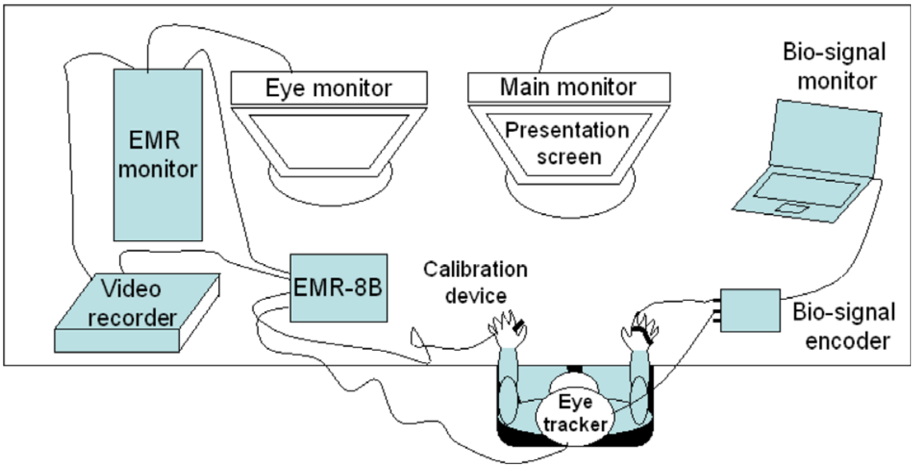

Fig. 4. Experimental setup.

and introducing the living room. In order to be able to compare the three versions, scenes where the agent or text box move 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 map or layout area is the field on the screen that displays the layout (map) of the room. The layout area is a designated reference object with fixed location. Depending on the room currently presented, the respective field in the layout is highlighted. Observe that in Fig. 2, the location of the living room is highlighted.
- Other screen areas.

3.7. 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, and will be discussed in this section. Spatio-temporal analysis will be discussed in the following section.

3.7.1. Focus of attention hypothesis

In order to test the Focus of Attention Hypothesis 1, we specifically investigated 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 referring entity (agent, text and voice) refers to some item of the apartment. A 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 ($F(2,9) = 8.2; p < .01$). (The level of statistical significance is set to 5%.) The independent variable refers to one of the three conditions, whereas the dependent variable is the number of gaze points in one of the designated areas. 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 ($F(2,9) = 2.8; p = .11$).

The 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, the attentional focus of subjects is more evenly shared between the presentation medium and the presented material. While this result can hardly be seen as surprising, it can not be assumed a priori.

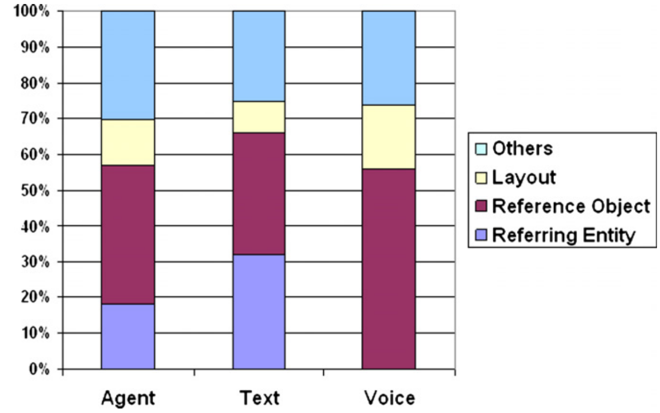

Fig. 5. Impact of Agent vs. Text vs. Voice version on the number of gaze points in different screen areas.

3.7.2. Locked attention hypothesis

Hypothesis 2 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 4429.3 gaze points (stdev = 2050.4), which corresponds to 18% of the total number of gaze points, and the mean for the text box is 7600.8 points (stdev = 2350.8), amounting to 32% of the total number of gaze points (see Fig. 5). The *t*-test (one-tailed, assuming unequal variances) showed that subjects look significantly more often at the text box than at the agent ($t(6) = -2.47; p < .05$).

This result can be seen as evidence that subjects spend considerable time for processing the text in the box. It can be explained by the fact that new information is gradually revealed and by the nature of the stimulus itself (textual information), which requires additional processing time. The issue is that locked attention can prevent users from attending to other salient information (Faraday and Sutcliffe, 1996).

3.7.3. Agent face–body hypothesis

The Agent Face–Body Hypothesis 3 has been tested by summarizing gaze points that are contained in either the agent face or the agent body region. The mean was 3322.2 gaze points (stdev = 634.6) for the face, and 727.4 (stdev = 455.6) for the body. It could be shown that subjects were looking mostly at the agent's face (mean = 83.1), which can be interpreted as supportive evidence for the hypothesis that users interact socially with life-like interface agents (see Witkowski et al. (2001) for a similar result).

This result can be regarded as related to the CASA (Computer As Social Actor) effect, demonstrating that people apply social rules in human–computer interaction in ways that are similar to human social interactions (Reeves and Nass, 1998). Since the CASA effect merely assumes a computer with anthropomorphic voice, the presence of a visual life-like character seemingly transfers to social etiquette, such as looking into the face of one's interlocutor. Another interpretation of the result is that subjects were simply attending to the apparent source of the voice sound, the moving mouth of the agent.

3.8. 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 *spatio-temporal* analysis of eye movement data with 22 sentences, which involves the investigation of the temporal evolution of users' attention to different interface objects. In the following, we present our observations.

3.8.1. Extended Auditory Language Processing Hypothesis

We first discuss the Extended Auditory Language Processing Hypothesis 4 with respect to our three conditions. In Fig. 6, the referring entity (agent, text box and 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 words in the study in Cooper (1974), 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).

The sentence in Fig. 7 is similar to the sentences used in Cooper (1974) as it contains a 'trigger word' – here the word "window" that is both spoken and has a unique semantically related visualization (the picture of a window). In the Agent version two subjects focus on the visual window shortly after they hear the word "window" (as predicted by Cooper (1974)). However, the two subjects were already looking at the window before the word "window" is uttered. A likely reason is that the agent performs a deictic (facial) gesture turning its head to the direction of the window shortly before uttering the sentence. The Voice version does not show a clear attentional pattern of subjects' eye movements. In line with the Locked Attention Hypothesis 2, 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.

3.8.2. Social Interaction Protocol Hypothesis

As a first attempt to provide a systematic spatio-temporal analysis of eye movements for interfaces with visual navigational aids, we propose an Instructor–Reference–Instructor (IRI) triple as a basic unit for evaluating the Social Interaction Protocol Hypothesis 5. An IRI denotes a situation where the user first attends to an instructor,

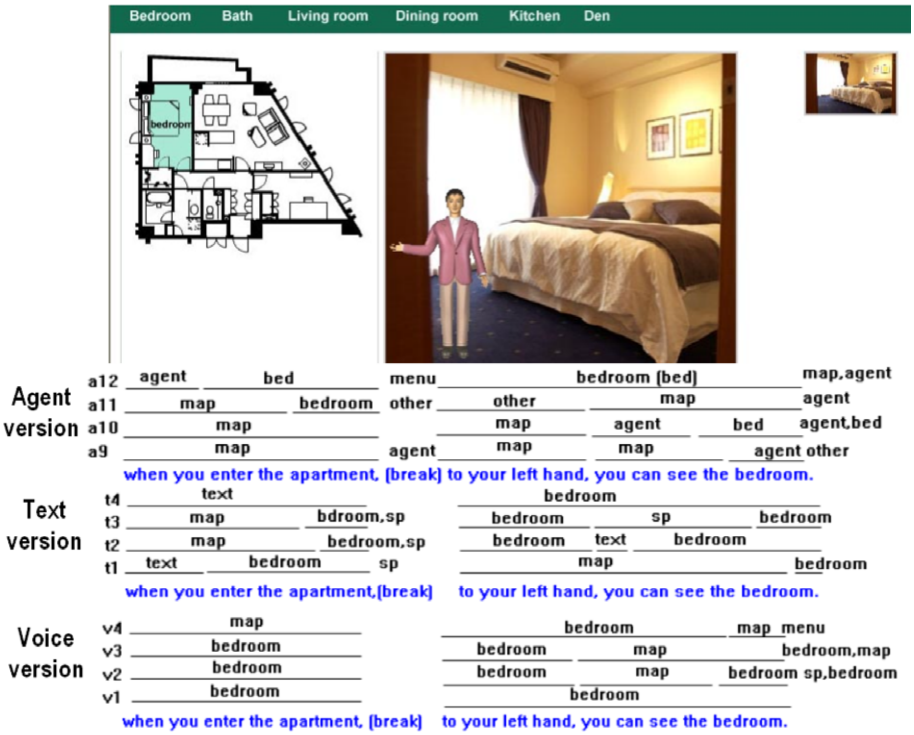
Bedroom
Bath
Living room
Dining room
Kitchen
Den
bedroom
Agent version
a12 agent bed menu bedroom (bed) map,agent
a11 map bedroom other other map agent
a10 map map agent bed agent,bed
a9 map agent map map agent other
when you enter the apartment. [break] to your left hand, you can see the bedroom.
Text version
t4 text bedroom
t3 map bdroom,sp bedroom sp bedroom
t2 map bedroom,sp bedroom text bedroom
t1 text bedroom sp map bedroom
when you enter the apartment,[break] to your left hand, you can see the bedroom.
Voice version
v4 map bedroom map menu
v3 bedroom bedroom map bedroom,map
v2 bedroom bedroom map bedroom sp,bedroom
v1 bedroom bedroom
when you enter the apartment. [break] to your left hand, you can see the bedroom.

Fig. 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,... at designated interface objects. (The abbreviation "sp" refers to the small picture to the top-right.)

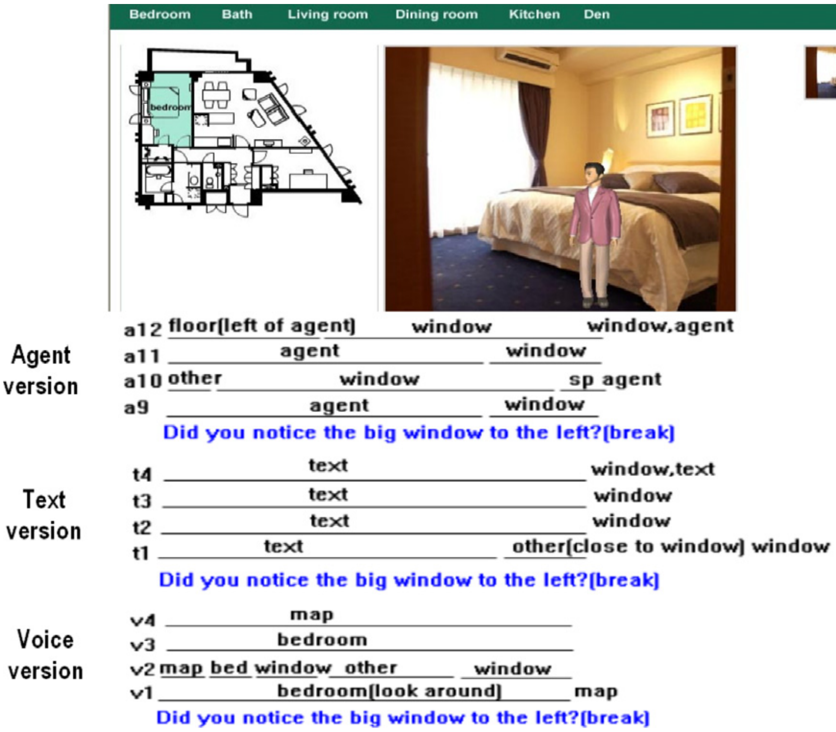
Bedroom
Bath
Living room
Dining room
Kitchen
Den
bedroom
Agent version
a12 floor(left of agent) window window,agent
a11 agent window
a10 other window sp agent
a9 agent window
Did you notice the big window to the left?(break)
Text version
t4 text window,text
t3 text window
t2 text window
t1 text other(close to window) window
Did you notice the big window to the left?(break)
Voice version
v4 map
v3 bedroom
v2 map bed window other window
v1 bedroom(look around) map
Did you notice the big window to the left?(break)

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

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 (Nakano et al., 2003), and indicators of the instructor being conceived of as a social actor.

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, and eventually fixate on the layout.

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 table shows that in the Agent version, subjects look back to the instructor at sentence breaks more than in the Text version, and tend to more accurately shift their attention to the intended reference object.

Table 1
Shift of attentional focus (i) to agent or text box at sentence breaks and (ii) to intended object of reference by referential acts of the agent or text box

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%

The attentional shifts suggest that subjects can perceive animated agents to possess a certain degree of competence, such as being competent in directing the user to locations of interest. Moreover, 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.

However, given that even infants tend to follow their mother's eye gaze or direction of pointing without necessarily attributing intentions or mental states, we should be careful in over-interpreting the reason for the subjects' eye response. Although an 'Arrow version' has not been tested in the current study, it is likely that subjects would also follow the directions of an arrow (Takeuchi and Naito, 1995).

3.8.3. Cost of Media Hypothesis

In order to investigate subjects' overall state of arousal or stress during the presentation (Hypothesis 6), their bio-signals were analyzed. Since the signals values of subjects may vary significantly depending on individual differences, room temperature, and other factors, physiological values were first normalized by applying the operation $(\mathrm{AM}_{\mathrm{pres}} - \mathrm{AM}_{\mathrm{relax}})/Range$, whereby $\mathrm{AM}_{\mathrm{pres}}$ and $\mathrm{AM}_{\mathrm{relax}}$ are the means of the presentation and baseline periods, respectively, and *Range* is defined as $x_{\max} - x_{\min}$ for signal x. Intuitively, a smaller value indicates that interaction with the interfaces has a (overall) more calming effect on the user (derived from skin conductance) or decreases negative feelings (derived from heart rate) to a higher extent.

The operation implements an approximation to assessing subjects' affective state in that signal values are summarized for the whole presentation period rather than for designated partitions of the presentation. However, unlike our previous studies (Prendinger and Ishizuka, 2005; Prendinger et al., 2005b), the presentation of the apartment does not obviously feature segments in which particular emotions would be elicited. Normalized values for each condition (agent, text and voice) where calculated but no significant differences were found. Also, no significant differences were obtained for HR.

In summary, the study did not support the hypothesis that presentations guided by different media, such as an agent, a text box, or speech only, lead to significantly different physiological signal levels.

3.9. Questionnaire results

In addition to eye and biometric 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 one on the subjects' ability to recall shown items.

In the first set of questions, subjects were asked:

(Q1) Whether they would want to live in the apartment;
(Q2) Whether they would recommend the apartment to a friend; and
(Q3) 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 (Q1) and (Q2) 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 (Q3), however, showed that subjects judged the Voice version to be more helpful than either of the other versions ($F(2,12) = 8.9$; $p < .01$). 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 and recall. The latter result might be explained by the fact that all but one question were related to room items that were not mentioned in the verbal comments. Since subjects in the Voice version were not distracted by the agent or text box, they had more time to scan the rooms, and might therefore have better remembered shown items.

4. Discussion

This paper has introduced a novel method for evaluating the utility of life-like interface agents, which is based on tracking users' eye movements, an objective evaluation method that does not distract the user from the primary task. Although eye tracking has been abundantly used in psychology, multimedia, and related studies (Duchowski, 2003), 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 show that life-like interface agents have a noticeable effect and may provide a more social interface to online information. Users follow the verbal and non-verbal navigational directives of the agent and mostly look at the agent's face. However, the latter mentioned result begs the question whether subjects were aware of the agent's deictic 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) observations while looking at the videos. When the agent performs a deictic 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, or follows the agent's gaze direction.

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. These results also forward the discussion about the believability of life-like agents and the 'Characters As Social Actors' effect 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 direct the user's focus of attention to objects of interest while maintaining aspects of the social interaction protocol. However, we certainly have to be careful not to over-interpret the results of a pilot study with small sample size. We also have to be cautious when interpreting the social effect of interface agents. It might well be true that different kinds of animated interface object attract users' attention, even non-anthropomorphic ones, such as an arrow.

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 Witkowski et al. (2001), our work is the first one 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 (Bates, 1994), and hence, worth investigating.

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 Wilson and Sasse (2000), 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 van Mulken et al. (1998)). However, it should be emphasized that unlike the work of Moreno (2004), improving student learning with life-like agents was not the rationale for this 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 (Cassell et al., 2000). 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 results of the current study can be seen as support for the claim 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 Goldberg and Kotval (1999) 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 (Umemuro and Yamashita, 2003) and for affective interest (Hess, 1972; Partala and Surakka, 2003).

Another 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. Nakano et al. (2003) 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’.

In terms of the future of interfaces employing life-like interface agents, the study in this paper was 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 Prendinger and Ishizuka (2005) 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 will greatly advance interfaces that realize 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, 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), by a JSPS Encouragement of Young Scientists Grant (FY2005–FY2007), and a NII Joint Research Grant with the University of Tokyo (FY2005).

References

André, E., Müller, J., Rist, T., 1996. The PPP Persona: A multipurpose animated presentation agent. In: Proceedings Advanced Visual Interfaces (AVI-96). ACM Press, New York, pp. 245–247.

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

Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.), 2000. Embodied Conversational Agents. The MIT Press, Cambridge, MA.

Cooper, R.M., 1974. 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.

Craig, S.D., Gholson, B., Driscoll, D.M., 2002. Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. Educational Psychology 94 (2), 428–434.

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

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

Faraday, P., Sutcliffe, A., 1996. An empirical study of attending and comprehending multimedia presentations. In: Proceedings of ACM Multimedia 96, Boston MA, pp. 265–275.

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

Hess, E.H., 1972. Pupillometrics: A method of studying mental, emotional and sensory processes. In: Greenfield, N., Sternbach, R. (Eds.), Handbook of Psychophysiology. Holt, Rinehart & Winston, New York, pp. 491–531.

Hongpaisanwiwat, C., Lewis, M., 2003. Attention effect of animated character. In: Proceedings Human–Computer Interaction (INTERACT-03). IOS Press, pp. 423–430.

Hottolink, 2002. Hottolink Inc. Available from: http://www.hottolink.com.

Jacob, R.J.K., 1991. 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.

Klein, J., Moon, Y., Picard, R., 2002. This computer responds to user frustration: Theory, design, and results. Interacting with Computers 14, 119–140.

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

Levenson, R.W., 2003. Autonomic specificity and emotion. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (Eds.), Handbook of Affective Sciences. Oxford University Press, Oxford, pp. 212–224.

Lin, Y., Zhang, W.J., Koubeck, R.J., 2004. Effective attention allocation behavior and its measurement: A preliminary study. Interacting with Computers 16, 1195–1210.

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

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

Moreno, R., 2004. Animated pedagogical agents in educational technology. Educational Technology 44 (6), 23–30.

NAC, 2004. Image Technology. Available from: http://eyemark.jp.

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

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

Oyekoya, O.K., Stentiford, F.W.M., 2004. Eye tracking as a new interface for image retrieval. British Telecommunications Technology Journal 22 (3).

Partala, T., Surakka, V., 2003. Pupil size variation as an indication of affective processing. International Journal of Human–Computer Studies 59, 185–198.

Picard, R.W., 1997. Affective Computing. The MIT Press, Cambridge, MA.

Prendinger, H., Ishizuka, M. (Eds.), 2004. Life-Like Characters. Tools, Affective Functions, and Applications. Cognitive Technologies. Springer Verlag, Berlin, Heidelberg.

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

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

Prendinger, H., Ma, C., Yingzi, J., Nakasone, A., Ishizuka, M., 2005a. Understanding the effect of life-like interface agents through eye users' eye movements. In: Proceedings of Seventh International Conference on Multimodal Interfaces (ICMI-05). ACM Press, New York, pp. 108–115.

Prendinger, H., Mori, J., Ishizuka, M., 2005b. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. International Journal of Human–Computer Studies 62 (2), 231–245.

Qu, L., Wang, N., Johnson, W.L., 2004. Pedagogical agents that interact with learners. In: AAMAS-04 Workshop on Balanced Perception and Action in ECAs, New York, NY, USA.

Reeves, B., Nass, C., 1998. The media equation. How People Treat Computers, Television and New Media Like Real People and Places. CSLI Publications, Center for the Study of Language and Information. Cambridge University Press.

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

Rist, T., André, E., Baldes, S., Gebhard, P., Klesen, M., Kipp, M., Rist, P., Schmitt, M., 2004. A review of the development of embodied presentation agents and their application fields. In: Prendinger, H., Ishizuka, M. (Eds.), Life-like Characters. Tools, Affective Functions and Applications. Cognitive Technologies. Springer, Berlin, Heidelberg, pp. 377–404.

Salvucci, D.D., Goldberg, J.H., 2000. Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the Eye Tracking Research and Applications Symposium. ACM Press, New York, pp. 71–78.

Takeuchi, A., Naito, T., 1995. Situated facial displays: Towards social interaction. In: Proceedings CHI 95 Conference. ACM Press, New York, pp. 450–455.

Tokyo Mansions, 2004. Tokyo Mansions. Available from: http://www.themansions.jp/.

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

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

Ward, R., Marsden, P., 2003. Physiological responses to different WEB page designs. International Journal of Human–Computer Studies 59, 199–212.

Wilson, G., Sasse, M., 2000. Listen to your heart rate: Counting the cost of media quality. In: Paiva, A. (Ed.), Affective Interactions – Towards a New Generation of Computer Interfaces. Springer, Berlin, Heidelberg, pp. 9–20.

Witkowski, M., Arafa, Y., de Bruijn, O., 2001. Evaluating user reaction to character agent mediated displays using eye-tracking technology. In: Proceedings AISB-01 Symposium on Information Agents for Electronic Commerce. pp. 79–87.