Attentive presentation agents

Tobias Eichner, Helmut Prendinger, Elisabeth André, M. Ishizuka

Angaben zur Veröffentlichung / Publication details:

Eichner, Tobias, Helmut Prendinger, Elisabeth André, and M. Ishizuka. 2007. "Attentive presentation agents." In *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007*, edited by Catherine Pelachaud, Jean-Claude Martin, Elisabeth André, Gérard Chollet, Kostas Karpouzis, and Danielle Pelé, 283–95. Berlin [u.a.]: Springer. https://doi.org/10.1007/978-3-540-74997-4_26.

Attentive Presentation Agents

Tobias Eichner[1,2], Helmut Prendinger[1], Elisabeth André[2],
and Mitsuru Ishizuka[3]

[1] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,Tokyo 101-8430, Japan
helmut@nii.ac.jp
[2] Institute of Computer Science, University of Augsburg
Eichleitnerstr.30, D-86135Augsburg, Germany
tobias.eichner@gmail.com, andre@informatik.uni-augsburg.de
[3] Graduate School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
ishizuka@i.u-tokyo.ac.jp

Abstract. The paper describes an infotainment application where life-like characters present two MP3 players in a virtual showroom. The key feature of the system is that the presenter agents analyze the user's gaze-behavior in real-time and may thus adapt the presentation flow accordingly. In particular, a user's (non-)interest in interface objects and also preference in decision situations is estimated automatically by just using eye gaze as input modality. A formal study was conducted that compared two versions of the application. Results indicate that attentive presentation agents support successful grounding of deictic agent gestures and natural gaze behavior.

1 Introduction

Advances in multi-modal user interfaces allows us to develop exciting new applications. Keyboard and mouse are no longer the sole input devices for interacting with computers. More subtle and natural ways of interaction can be supported and hence enrich the experience of the user interacting with an application.

In our work we propose a so-called infotainment (information and entertainment) application in which two life-like characters present two MP3 players of a fictitious company as co-presenters. The key feature of the system is that the agents are 'aware' of the user seated in front of the screen, which is achieved by the use of a video-based eye tracker. The system analyzes the gaze behavior of users in real-time and is capable of adapting the presentation flow according to the user's interest or non-interest. The system aims to imitate human presenters who typically receive (visual) feedback from their audience. If, e.g., a museum guide notices that people are distracted or interested in an object different from the one currently presented, the human guide can either try to regain the attention or shift the focus of the presentation to the other object.

This paper is structured as follows. Section 2 discusses related work. Section 3 sketches the MPML3D and Java Real-Time Component. Section 4 describes our methods to recognize (visual) interest and preference. Section 5 provides details about the application scenario, available agent responses, and our notion of 'grounding'. In Section 6 we put forth hypotheses about attentive presentation agents. Section 7 describes our empirical study using two versions of the application. The results are presented in Section 8. Section 9 concludes the paper.

2 Related Work

One of the early uses of human gaze in human–computer interaction was as a pointing device [8]. Similar to positioning a mouse pointer on an interface object in order to activate it, a user might trigger an action in the interface by simply looking at the object. Rather than using gaze for controlling or manipulating interface objects, Attentive User Interfaces (AUIs) [21] and "visual attentive interfaces" [16], by contrast, are geared toward recognizing the user's intention from natural gaze behavior. For instance, the InVision system described in [16] processes a user's gaze directed at an interface depicting a kitchen environment, and infers whether the user is hungry or intending to rearrange the kitchen items, and so on, from the gaze path.

The "gaze-responsive self-disclosing display" [19] is one of the first systems which analyzes the user's gaze behavior in real-time and responds to it in an appropriate way. The application shows a small 3D planet, where a 2D virtual agent tells something about the items one can see on the planet. When the user looks on some specific object, e.g. a staircase, the virtual narrator will talk about the staircase. If the user's gaze switches between two staircases, the agent will talk about staircases as a group. Additionally, the current object of interest is zoomed-in. The FRED system [20] makes use of 3D animated facial agents and combines them with a conversational gaze model in a multi-agent setting. The agents have the capability to notice if the user (or another agent) is looking at them. Together with the speech data they can determine if they have to listen to someone else or if they can talk. The focus of that work is the regulation of conversional flow in a multi-agent environment, whereas we focus on detecting interest and attentiveness in a virtual presentation.

Another application using an virtual agent is the MACK system described in [10]. The authors uses a head tracker to determine a user's gaze in a direction-giving task. The animated agent explains directions on a map and monitors the user's head. In that application, lack of negative feedback indicates successful grounding. If grounding fails, the agent will perform a repair action to help the user. The difference of our work to the MACK system is that we do not assume verbal input to drive the presentation of the agent. Furthermore, we analyze and interpret eye movements rather than head movements.

3 MPML3D and Java Real-Time Component

Our application is based on (1) the MPML3D framework for animating the agents and defining the content of the presentation, (2) a Java Real Time

Component for receiving and analyzing the eye data in real-time, and (3) the eye-tracking system. (1) and (2) will be introduced in this section, and (3) in Sect. 7.

3.1 MPML3D Framework

When setting up a virtual presentation, the focus of the author should be on creating content and not on technical issues. The Multimodal Presentation Markup Language for 3D agents (MPML3D) provides a easy-to-use XML scripting language to define both the content of the presentation as well as the animations of the agents [7,11]. The main concept of MPML3D are *actions*: every (speech) utterance of an agent, every gesture and mimic expression, etc. is defined as an action. Actions can be performed sequentially or in parallel. Hence it is possible that an agent performs, e.g. a deictic gesture during an utterance. Actions can be broken down into sub-actions, which allows synchronization word by word.

An important feature of actions is that they can be interrupted by other actions anytime and thus react to certain events triggered by user input instantly. An interruption can be induced by the concept called *perceptions*: unlike actions, they run in the background and wait for an event to occur. Perceptions can be defined for arbitrary objects. So it is possible that the agents perceive themselves or, e.g., which screen object the user is attending to. Perceptions can be seen as a core feature of MPML3D by which agents sense their virtual environment and the real world (using sensors).

The main difference between MPML3D and other character markup languages, such as the Affective Presentation Markup Language (APML) [2] is that MPML3D focusses on the control and interaction between two (or more) agents rather than a single agent.

3.2 Java Real Time Component

The second basic component of our application is the Java Real-Time Component (JRTC). An eye tracker provides huge amounts of data during tracking, amongst others gaze coordinates, pupil size, eye closure, direction vector of the head, etc. In order to access all these data in an easy and convenient way, we have developed the JRTC component. It provides a form of API, making it simple to integrate the data of the eye tracker into any application. At the moment, JRTC supports the Seeing Machines eye tracker [15] and a device for bio-signal processing. JRTC not only receives data from devices, but also performs computations on the data. It can deliver both raw data and classification results or smoothed data. To be even more flexible and powerful, a Bayesian Net can be included with the Netica library [12]. In our application, we only made use of the gaze coordinates. All the computations (introduced in the next section) were also integrated in the JRTC.

4 Gaze-Based Recognition of Interest and Preference

A basic functionality of our presentation system is to recognize the visual interest of the user, i.e. which interface object the user pays attention to. Another, independent capability of the system is to determine the preferred visual object among two by only analyzing the gaze behavior. Both were introduced in [6].

In our system, *visual interest* was determined by a slightly simplified version of the algorithm introduced in [14], where two interest metrics have been introduced: The *Interest Score* (IScore) metric indicates the 'arousal' of a visual stimuli, i.e. the probability that the user is interested in that visual object. When the IScore value passes a certain threshold, the object becomes 'active'. The *Focus of Interest Score* (FIScore) calculates the amount of interest in an active object over time. If the FIScore for an active object falls below a certain threshold, it becomes deactivated (the user lost interest in that object).

The main component to calculate the IScore metric is $p = \frac{T_{ISon}}{T_{IS}}$, where T_{ISon} is the accumulated time on a visual object within a sliding time window of the size T_{IS} (set to 1000 ms). To allow for other factors influencing the calculation of interest, [14] suggest the following, extended equation: $p_{is} = p(1+\alpha(1-p))$. Here, α represents a set of parameters increasing the accuracy of interest estimation. Our simplified version has two out of the four parameters defined in [14]:

$$\alpha = c_0 \frac{c_f \alpha_f + c_s \alpha_s}{c_f + c_s}$$

α_f represents the frequency of entering and leaving an object with the gaze. The more often the user's gaze switches to an object, the higher is the chance of excitation. It is calculated with the formula $\alpha_f = \frac{N_{sw}}{N_f}$, where N_{sw} denotes the number of times eye gaze enters and leaves the object and N_f denotes the maximum possible N_{sw} in the preset time window. α_s is calculated as $\alpha_s = \frac{S_b - S}{S}$, where S_b denotes the average size of all defined interest objects and S the size of the currently calculated object. It is assumed that larger visual objects have a higher chance to be hit by chance than smaller objects because of noise in the eye data. The factor compensates for this issue.

FIScore calculates the continued interest of the user in the active object over time. Like in IScore, the basic component is the gaze intensity on the active object. Additional, gaze intensity on other interest objects is considered. The sliding time window is twice as big as for IScore computation (2000 ms).

To estimate *(visual) preference*, we exploited the so-called "gaze cascade" effect in two-alternative forced choice (2AFC) situations. This effect was discovered in a study where users had to choose the more attractive face from two faces [17]. It could be demonstrated that there was a distinct gaze bias towards the chosen stimulus in the last one and a half seconds before the decision was made. The decision formation process was completed within six to seven seconds. Based on these results, the AutoSelect was developed for real-time estimation of preference [1]. Thus gaze points are calculated in a time window of 1500 ms length. If 90% (or more) of all gaze points within this time window are on one visual object, the system chooses that object. AutoSelect was tested in a study where users had to select their preferred necktie from two presented neckties by gaze. The system achieved an accuracy of 72%.

5 Gaze-Based Presentation

Our application consists of a virtual sales scenario where two MP3 players are presented by a team of a female and a male agent, which was first described in [13]. The models of our agents were designed by a professional Japanese artist. They can perform several gestures like greeting, counting with fingers, deictic and beat gestures, and facial expressions (happy, sad, surprised). Speech output is generated using Loquendo Text-to-Speech [9], and lips are adequately animated.

The course of the presentation can be summarized as follows: Yuuki, the female agent, starts the presentation by introducing the (fictitious) company and her colleague Ken (the male agent). After that, Ken promotes the first MP3 player, the MP3PodAdvance, by providing a description of its features, which includes an example of navigating the menu of the player to select a particular song. After Yuuki presents the other MP3 player, the EasyMP3Pod, both agents argue over the benefits and drawbacks of each player. During that discussion they realize that the device presented by the other character would fulfill their particular needs better than what they have presented themselves. Hence, the agents address the user directly, and ask him or her to choose one of the two MP3 players. Finally, the two agents say good bye and the presentation ends.

5.1 Reacting to the User's Interest State

The key feature of the application is that the user experiences that the virtual agents are aware of him or her and react to interest and non-interest. The user should not have the impression of watching a static presentation. To achieve this goal, we defined gaze-sensitive areas on the screen, so-called "interest objects". In these areas, the user's gaze is analyzed regarding (non-interest in the presentation. The defined objects are (from the left to the right): (i) SideAds, a total of four slides that advertise the MP3 players and are exchanged every five seconds; (ii) male agent ("Ken"); (iii) 3D model of MP3PodAdvance; (iv) virtual slide; (v) 3D model of EasyMP3Pod; (vi) female agent ("Yuuki"); (vii) the view out of the window. Their bounding rectangles can be seen in Fig. 1.

For the following four screen objects, agent reactions are defined when the user shows interest in them.

SideAds: Motivated by the experimental design (described in Sect. 7), the changing side ads should distract the user. If the user looks at the ads for the second time, the currently not speaking agent will interrupt the presentation asserting that the user is distracted. Afterwards, the side ads are turned off.

Ken and Yuuki: Both agents know when the user is looking at them. There will be an interruption if the user is looking at the non-speaking agent. E.g., Ken

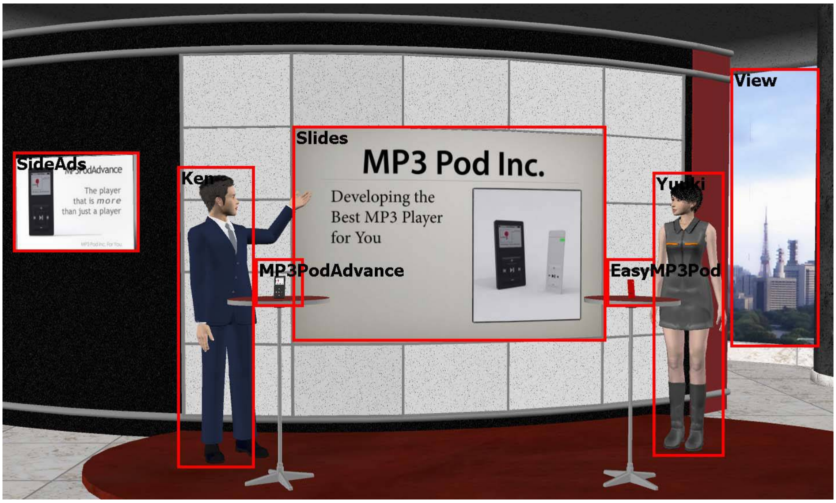

Fig. 1. Bounding areas of interest objects. (The displayed computer screen area is clipped for convenience and does not show the view the user has).

will state "Is there something wrong with my necktie? I think you are looking at me, even if Yuuki is explaining something."

View: If the user shows interest in the view out of the window over Tokyo, the agents will interrupt the presentation, the camera of the 3D scene will focus the skyline and Yuuki will talk about some landmarks. This is the only exception where the camera focuses an interest object.

Off-screen: As a sign of general non-interest can be seen when the user takes the gaze off the screen area. In that case, the agents wonder what they can to to regain the user's attention. After a short chat they continue the presentation.

Notice that these interruptions can occur anytime during the presentation depending on the user's gaze behavior, i.e. they are context-independent. In the following, we will describe context-dependent interruptions.

5.2 Responding to Failed Grounding

An important indicator of a user's attentiveness is successful grounding. In human face-to-face communication, grounding relates to the process of ensuring that what has been said is understood by the conversational partners, i.e., there is "common ground" [4,3]. During the presentation, the agents repeatedly perform referential gestures to link their spoken content to a visual stimulus. Our system checks grounding in these situations as follows:

- If the grounding situation lasts for less than 2000 ms, it is treated as a *short grounding situation*. In this case, the user is supposed to look at the grounding object (the referred visual stimuli) during the utterance or within

one second after the utterance or gesture terminated for at least 150 ms. This situation typically occurs when the agents perform deictic gestures or explain a changing region on the virtual slides.
- In *long grounding situations* (longer than two seconds), the user is supposed to look at the grounding object for 45% of the time of the duration of the utterance. This situation occurs when an agent explains facts on the (virtual) slides with no changing content.

In our application, the two agents, the two virtual models of the MP3 players and the slides are possible grounding objects. If the system classifies a grounding situation as failed, the agents will react on this. In total, 15 grounding situations are defined, whereby the slides are defined as grounding object in nine cases, the two player models in two cases each, and Ken and Yuuki in one case each.

Here are some examples of available reactions:

Yuuki's self-introduction: At the beginning of the presentation, Yuuki introduces herself. The user is assumed to look at her during this time. If not, Ken will interrupt and ask the user to look at Yuuki.

Ken's introduction: When Yuuki introduces Ken, the user's gaze should transit from Yuuki to Ken during the deictic gesture. Otherwise, Ken will wave his hand and 'help' the user to find him on the screen.

Slides: The majority of grounding situations is related to the slides. Here the agent explains the slide content, often accompanied by a deictic gesture. If the user does not attend to the slide, an agent will ask the user for more attention.

MP3 players: When the agents refer to the players by using a deictic gesture, the user's focus should turn to the model. The agents monitor whether the user attends to the player and react, if he or she is not looking there.

5.3 Preference Estimation

At the end of the presentation, the agents ask the user to choose the preferred player. For that purpose, images of both players are shown on the slide and the agents perform deictic gestures towards them (see Fig. 2). A pre-study showed that the gaze cascade phenomenon will occur naturally in this situation: Users will alternately look left and right on the slide and exhibit a bias for one player. The gaze-sensitive bounding boxes include both the images on the slide and the two models of the player in case that users consider the virtual models in the decision making process. At the end of the automatic decision time of the system (7.5 seconds), the user is asked

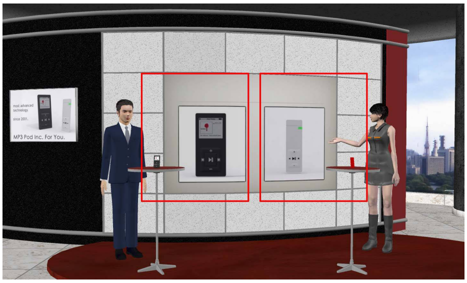

Fig. 2. Bounding areas for selecting the preferred MP3 player

to declare the decision by a key press on a keyboard. This allows us to compare the decision calculated by our system with the user's decision. Finally, the agent with the preferred player expresses happiness about its successful promotion.

6 Theory

The general hypothesis of our research on attentive presentation agents is that they can provide a more natural interaction experience such that the user will experience the presentation agents as more mindful and exhibit a more natural gaze behavior towards the presentation. We suggest the following hypotheses:

Grounding Hypothesis: Grounding is more successful with attentive agents.

Mindfulness Hypothesis: It is speculated that users experience the interaction with attentive agents as similar to human face-to-face communication and more engaging, inducing a sense of involvement and co-presence with the presentation agents. Our operationalization of 'mindfulness' is based on the concepts of "face-to-face"(communication), "involvement", "co-presence", and "partner evaluation" proposed in [5], and "engagement" as described in [18].

Gaze Cascade Hypothesis: Based on our previous finding [1], we predict that the gaze cascade effect generalizes to the virtual product presentation scenario.

7 Method

Experimental Design. The experiment had a between-subjects design. Two versions were implemented.

– *Interactive (I) version*: The system analyzes the user gaze and the agents react accordingly. Agent interruptions cannot be foreseen in this version.
– *Pseudo-interactive (PI) version*: The agents interrupt the presentation at seven pre-defined points, independent of user gaze. Since agents *do* react, even in a seemingly random way, this version is called "pseudo-interactive".

Note that all interruptions that could occur in the interactive version, actually do occur in the pseudo-interactive version. This allows us to compare the two versions, because there are no reactions which can only occur in one version.

Subjects. Thirty-five subjects, all students, researchers, or staff from NII, participated in the study. Their age ranged from 19 to 41 yrs (average 27.5 yrs). They received 1,000 Yen for participation. There were technical difficulties with ten subjects due to lacking experience of the experimenter with the eye tracking setup. Four subjects could not be calibrated because of reflections of contact lenses or glasses. All of those subjects were excluded from the study beforehand. The remaining twenty-one subjects (twelve female, nine male) were randomly assigned to the interactive and pseudo-interactive versions.

Apparatus. The presentation was shown on an IBM 20.1 inch screen with a resolution of 1600 × 1200 pixels and ran on a Dell workstation with dual-core

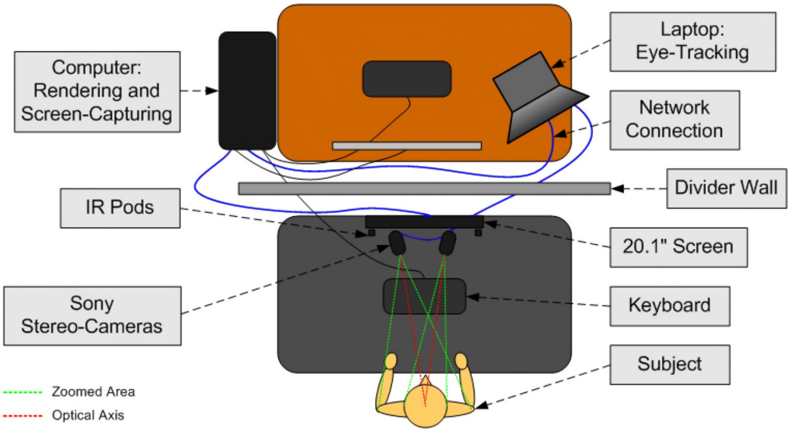

Fig. 3. Experimental system setup

processor. The eye-tracking software faceLAB from Seeing Machines [15] ran on a separate laptop which was connected to the workstation via network. Sony stereo cameras of the faceLAB eye tracker and loudspeakers were positioned below the screen. The user was seated in front of the screen (80 cm distance). Two infrared pods were attached at the upper part of the display for illuminating the eyes. The system has a sampling rate of 60 Hz. In real-time modus of faceLAB, data processing has a delay of 30 ms. Each presentation was captured as a video file. The schematic setup can be seen in Fig. 3.

Procedure. Subjects entered the experiment room individually and received a written instruction about the procedure. The instruction given to the subjects was to watch the presentation as they would watch a presentation given by human presenters. At that time, the experimenter was available for queries. Subsequently, each subject was calibrated for eye tracking. The subject was asked to assume a comfortable sitting position, and the experimenter started the calibration process by first determining reference points for head tracking, and then for eye (and pupil) tracking. Calibration is a step-by-step process following the menu-based instructions of the faceLAB software. The experimenter receives feedback on the accuracy of the calibration process, and may repeat some step, if necessary. Calibration of a subject took five minutes on average.

Then the subjects were shown the presentation, which lasted for about seven to eight minutes. During that time, only the experimenter and an assistant were present in the room and silence was kept. After the presentation finished, subjects were asked to fill out a questionnaire with nineteen questions that addressed their impression of the presentation.

8 Results

We start with some general results. The mean length of the presentation in the interactive version was 423.68 sec, and 477.13 sec in the pseudo-interactive

Table 1. Results of grounding success rate by version and category of grounding object

Grounding Category	I Version	PI Version
Yuuki	85.7%	100.0%
Ken	14.3%	50.0%
Slides	92.0%	76.0%
MP3 Players	53.5%	4.0%

version. One subject was removed from the pseudo-interactive version because of partly missing data. The interactive version was significantly shorter than the pseudo-interactive version ($t(18) = 2.33; p < 0.02$). These values confirm our prediction as interruptions in the interactive version were gaze-dependent. It is also reflected in the standard deviation (SD) of 70.1 seconds for the length of the interactive version. The interactive version had 5.7 interruptions per subject (vs. 7 in the pseudo-interactive version). It is worth mentioning that 31.6% of the interruptions were triggered by a failed grounding with respect to one of the virtual MP3 players. The 3D models might have been too small or the deictic gestures to them were not sufficiently precise. Another explanation might be their simplicity as graphical objects and short visual encoding time.

To test the *Grounding Hypothesis*, we counted successful grounding situations in both versions. In the interactive version, fifteen grounding situations were defined that may lead to an agent reaction in case of negative evidence (failed grounding). The same grounding situations are also present in the pseudo-interactive version (although agents will not react to failed grounding unless by coincidence). Overall, grounding was successful in 77% of the cases in the interactive version, and in 56.67% in the pseudo-interactive version. The detailed grounding success for defined grounding objects (summarized in Table 1) shows that success is depending on the grounding object. In the interactive version, the agents were able to lead the user's attention to the slides and also to the MP3 players. We do not have an explanation for the low success rate for groundings on Ken, but the comparatively high success rate in the pseudo-interactive version is delusive: shortly after Ken's introduction by Yuuki, Ken alerts the user to look at him. An explanation for the lower success rate of Yuuki can be that in the interactive version the presentation was startet by the user's gaze on the slides and not like in the pseudo-interactive version by keystroke of the experimenter. Hence, some users were maybe looking a little bit clueless at the beginning of the presentation where the grounding situation occured.

In order to test the *Mindfulness Hypothesis*, we relied on questionnaires as a standard evaluation method. A seven point Likert scale was used, ranging from "−3" (strongly disagree) to "3" (strongly agree), with "0" as the neutral attitude. Fifteen questions in the dimensions face-to-face, involvement, co-prescence and agent evaluation have been borrowed from [5], the engagement dimension was derived from the description in [18]. Interestingly, the only significant results relate to questions that address the salient feature of each version in a quite direct way. Subjects in the interactive version felt that the agents were aware of them to a significantly higher extent ($t(19) = 2.48; p < 0.05$), and in the pseudo-interactive version, they thought that the agents react to them in a strange way ($t(19) = -1.78; p < 0.05$). (All t-tests are one-tailed.) Moreover, we can interpret the result to the question "I had the impression that the agents cared about my interest." as a tendency that subjects consider their interest being taken into account better in the interactive than in the pseudo-interactive version ($t(19) = 1.38; p = 0.09$).

Finally, we turn to the results for the *Gaze Cascade Hypothesis*. In eight cases (out of twenty-one), the system could not make a decision within the 7.5 seconds for automatic preference estimation. These subjects might have needed more time to select their preferred MP3 player. For the remaining 13 decision situations, the system achieved an accuracy of 76.9%. Table 2 shows the results split up for the interactive and pseudo-interactive version. It is noticeable that the result in the pseudo-interactive version is much better. We can only speculate about the reason. Subjects in the interactive version might have noticed that the agents react to their gaze behavior and have a certain "anticipation". Consequently, subjects might have felt restricted and uncertain in how they 'should' look. By contrast, subjects in the pseudo-interactive version experienced ill-timed agent reactions and therefore might have felt less guided.

Table 2. Results of automatic preference estimation. ("n/a" means "not applicable")

System Decision	I Version	PI Version
correct	30.0%	63.6%
wrong	20.0%	27.3%
n/a	50.0%	9.1%

9 Conclusion

We described an agent-based presentation system that relies on eye gaze (1) to adapt the presentation to the user's interest, and (2) to react if the user is inattentive. For this purpose, we implemented appropriate agent feedback to guide the user or to ask him or her for more attentiveness. Gaze was also used to assess preference among two presented visual stimuli. Visual interest was estimated using a previously developed algorithm [14]. Preference estimation was based on our previously developed algorithm for analyzing the case cascade effect in real-time [1]. Furthermore, the system implemented an algorithm for testing successful grounding related to interface objects. This is achieved by the agents' deictic gestures, verbal comments, or a combination of both.

Our interface using attentive presentation agents is intended to provide a personalized experience. Users should have the feeling that their interest (or lack of interest) matters to the agents. User statements like "In the beginning I was bored. And then suddenly Yuuki waved her hand and asked for more attention." or "She [Yuuki] was yelling at me!" provide informal evidence that this aim was achieved. We conducted an empirical study where two versions of the system were compared: an interactive version analyzing the user's gaze behavior in real-time with appropriate agent reactions to interest/disinterest, and a pseudo-interactive version based on randomly assigned interruptions. It could be shown that in the interactive version, grounding was more successful, and that users in this version felt that the agents were aware of them. An open issue is to find an optimal level of agent attentiveness. The poor results regarding the gaze cascade effect indicate that gaze-based agents may carry a certain risk of 'overdoing' attentiveness. We will address this issue in our future research.

Our work has demonstrated the power of using eye gaze as input. The next step is to complement gaze with an additional modality, e.g. to recover from situations where the system could not classify the user's (non-)interest correctly. Another interesting issue might be to adopt the visual attention methods developed for our infotainment application to other scenarios like virtual worlds or interactive games.

Acknowledgments

The first author was supported by an Int'l Internship Grant from NII under a Memorandum of Understanding with the Faculty of Applied Informatics at the Univ. of Augsburg. The 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 an NII Joint Research Grant with the Univ. of Tokyo (FY2006).

References

1. Bee, N., Prendinger, H., Nakasone, A., André, E., Ishizuka, M.: AutoSelect: What You Want Is What You Get. Real-time processing of visual attention and affect. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Weber, M. (eds.) PIT 2006. LNCS (LNAI), vol. 4021, pp. 40–52. Springer, Heidelberg (2006)
2. Carolis, B.D., Pelauchaud, C., Poggi, I., Steedman, M.: APML: Mark-up language for communicative character expressions. In: Prendinger, H., Ishizuka, M. (eds.) Life-like Characters. Tools, Affective Functions and Applications, Cognitive Technologies, pp. 65–85. Springer, Heidelberg (2004)
3. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
4. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (eds.) Perspectives on Socially Shared Cognition, pp. 127–149. APA Books, Washington (1991)
5. Garau, M., Slater, M., Bee, S., Sasse, M.A.: The impact of eye gaze on communication using humanoid avatars. In: Proceedings SIGCHI Conference on Human Factors in Computing Systems (CHI-01), pp. 309–316. ACM Press, New York (2001)
6. Hoekstra, A., Prendinger, H., Bee, N., Heylen, D., Ishizuka, M.: Highly realistic 3d presentation agents with visual attention capability. In: Proceedings 7th International Symposium on Smart Graphics (SG-07). LNCS, vol. 4569, pp. 73–84. Springer, Heidelberg (2007)

7. Ishizuka, M., Prendinger, H.: Describing and generating multimodal contents featuring affective lifelike agents with MPML. New Generation Computing 24, 97–128 (2006)
8. Jacob, R.J.K.: 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)
9. Loquendo Vocal Technology and Services, URL (2006), http://www.loquendo.com
10. Nakano, Y.I., Reinstein, G., Stocky, T., Cassell, J.: Towards a model of face-to-face grounding. In: Proceedings of Association for Computational Linguistics (ACL-03), pp. 553–561 (2003)
11. Nischt, M., Prendinger, H., André, E., Ishizuka, M.: MPML3D: a reactive framework for the Multimodal Presentation Markup Language. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 218–229. Springer, Heidelberg (2006)
12. Norsys Software Corp. Netica, URL (2003), http://www.norsys.com
13. Prendinger, H., Eichner, T., André, E., Ishizuka, M.: Gaze-based infotainment agents. In: Proceedings ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE-07), pp. 87–90. ACM Press, New York (2007)
14. Qvarfordt, P., Zhai, S.: Conversing with the user based on eye-gaze patterns. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI-05), pp. 221–230. ACM Press, New York (2005)
15. Seeing Machines. Seeing Machines, URL (2005), http://www.seeingmachines.com/
16. Selker, T.: Visual attentive interfaces. BT Technology Journal 22(4), 146–150 (2004)
17. Shimojo, S., Simion, C., Shimojo, E., Scheier, C.: Gaze bias both reflects and influences preference. Nature Neuroscience 6(12), 1317–1322 (2003)
18. Sidner, C.L., Kidd, C.D., Lee, C., Lesh, N.: Where to look: A study in human–robot engagement. In: International Conference on Intelligent User Interfaces, pp. 78–84. ACM Press, New York (2004)
19. Starker, I., Bolt, R.A.: A gaze-responsive self-disclosing display. In: Proceedings CHI-90, pp. 3–9. ACM Press, New York (1990)
20. Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In: Proceedings of CHI-01, pp. 301–308. ACM Press, New York (2001)
21. Zhai, S.: What's in the eyes for attentive input. Communications of the ACM 46(3), 34–39 (2003)