Gaze-Based Infotainment Agents

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

Tobias Eichner
Elisabeth André
Institute of Computer Science
University of Augsburg
Eichleitnerstr. 30, D-86135
Augsburg, Germany
tobias.eichner@gmail.com

Mitsuru Ishizuka
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan
ishizuka@i.u-tokyo.ac.jp

ABSTRACT

We propose an infotainment presentation system that relies on eye gaze as an intuitive and unobtrusive input modality. The system analyzes eye movements in real-time to infer users' attention, visual interest, and preference regarding interface objects. The application consists of a virtual showroom where a team of two highly realistic 3D agents presents product items in an entertaining and attractive way. The presentation flow adapts to the user's attentiveness and interest, or lack thereof, and thus provides a more personalized and user-attentive experience of the presentation.

Categories and Subject Descriptors

H5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces—*input devices and strategies, interaction styles, theory and methods*; H5.1 [**Information Interfaces and Presentation (e.g., HCI)**]: Multimedia Information Systems—*animations*

General Terms

Human factors

Keywords

Multi-modal presentation, eye tracking, interest recognition, preference detection

1. INTRODUCTION

With the recent progress in multi-modal interfaces, exciting new types of interactive entertainment applications are being created, including virtual games, audience-guided movies, augmented sports, virtual travel guides and tutors [4]. Since multi-modal interfaces support both multi-modal input interpretation (audio, visual, haptic) and multi-modal output generation (speech, graphics, gesture, gaze), rich interaction experiences become possible.

ACE'07, June 13–15, 2007, Salzburg, Austria.

As an entertainment application, our focus is on multi-modal infotainment (information and entertainment), where life-like animated agents act in the role of virtual presenters that use their multi-modal expressiveness to convey information in a believable and entertaining way. This research field has already seen a significant development in available systems. Starting with single-agent presenters in the mid-1990s, the latest generation can manage interactive performances of multiple agents and analyze multi-modal input, such as verbal utterances combined with pointing gestures [10]. However, unless a user performs some purposeful communicative act, interface agents will remain ignorant about the user's current intention or interest. An inattentive interface may break the user's illusion of sharing an experience or the context with virtual agents.

In this paper, we propose an agent-based infotainment system that monitors and interprets eye movements in order to adapt the presentation flow according to the interest (or non-interest) of a user watching the presentation. Our gaze-contingent system aims to emulate the behavior of human presenters, who often glance at listeners to obtain feedback regarding their level of attention. If a human presenter notices that the audience is not looking at the currently described object (the referent), but is diverted by some other object, he or she will try to regain the attention to the referent or follow the interest shift of the audience.

An early implementation of a visual attentive interface is the 'gaze-responsive self-disclosing display' described in [15]. Here a simple facial agent will comment on visualizations of everyday items (such as a staircase) on a virtual planet, if the user's interest in some item can be inferred from gaze. Our proposed system revives the 'self-disclosing display' concept and extends it in interesting ways. First, we use full-body three-dimensional (3D) life-like agents rather than a simplistic face. Using deictic arm gestures, those agents can perform grounding references [3] in the 3D environment, e.g. by pointing to a (virtual slide), and alert the listener if positive evidence is missing (the user is not looking at the slide). Besides interest (non-interest) detection, we also implemented an automatic preference estimation algorithm. In this way, we can determine the user's choice between two (visualized) alternatives.

The paper is structured as follows. Section 2 discusses related work. Section 3 describes our methods to assess (visual) interest and preference. Section 4 provides details about the eye tracking setup, application scenario, and available gaze-based agent responses. Section 5 discusses the pos-

sible benefits of gaze-based infotainment systems. Section 6 concludes the paper.

2. RELATED WORK

Gaze-based systems share the interaction principle with a particular type of noncommand interface, so-called "interest and emotion sensitive" (IES) media [16]. In an IES media system, eye gaze is measured to determine the subsequent branch of a multiplex script board that a user is likely interested in. It is 'emotion sensitive' in that it also analyzes a user's pupil dilation and blink rate, from which affective states (e.g. arousal) can be inferred. Attentive User Interfaces (AUIs) [18] and 'visual attentive interfaces' [12] consider gaze for two purposes: (1) as a control device to support (provide a context for) pointing with a cursor, and (2) as an input modality to detect a user's intention. For instance, in the kitchen InVision project reported in [12], gaze patterns are analyzed to understand whether a user is hungry, considering to arrange the environment, or thinking something else. Similarly, our system exploits natural eye movements for interest detection in a noncommand fashion.

Life-like agents are virtual characters that are designed to convey the illusion of life or 'suspend disbelief' [1], so that users interacting with them will apply social interaction protocols naturally, e.g. by attending and responding to them as they would to other humans [7]. With the notable exception of [15], researches on visual attentive agents are rare. The eye-based FRED system endows animated facial agents with a conversational gaze model in a multi-agent setting [17]. The agents can adjust their gaze direction depending on which agent the user looks at. The MACK system described in [5] uses a head tracker to determine a user's gaze in a direction-giving task, whereby an animated agent monitors lack of negative feedback and positive feedback in the grounding process. If grounding fails, the agent will react with an appropriate repair action. The difference of our work to the MACK system (as well as to the human–robot interaction in [14]) is that do not assume verbal input.

3. ESTIMATION OF INTEREST AND PREFERENCE

To determine focus of interest, we modified the algorithm described in [9], where it is used for an intelligent virtual tourist information environment (iTourist). Two interest metrics were developed in [9]: (1) the Interest Score (IScore) and (2) the Focus of Interest Score (FIScore). IScore refers to an object's 'arousal' level, i.e. the likelihood that the user is interested in that (visual) object. When the IScore metric passes a certain threshold, the object is said to become 'active'. The FIScore calculates the amount of interest in an active object over time.

For our purpose, a simplified version of the IScore metric was sufficient, as we only need to know whether a user's attention is currently on a particular screen area, or not. The basic component for IScore is $p = T_{ISon}/T_{IS}$, where T_{ISon} refers to the accumulated gaze duration within a time window of size T_{IS} (here, 1000 ms). In order to account for factors that may enhance or inhibit interest, [9] characterize the IScore as $p_{is} = p(1+\alpha(1-p))$. Here, α encodes a set of parameters that increase the accuracy of interest estimation. We used two parameters: (1) the frequency of the user's eye gaze 'entering' and 'leaving' an object (α_f), and (2) the average size of all possible interest objects compared to the size of the currently computed object (α_s), which is intended to compensate for differences in the size of potential interest objects, and the related difference of being 'hit' by chance.

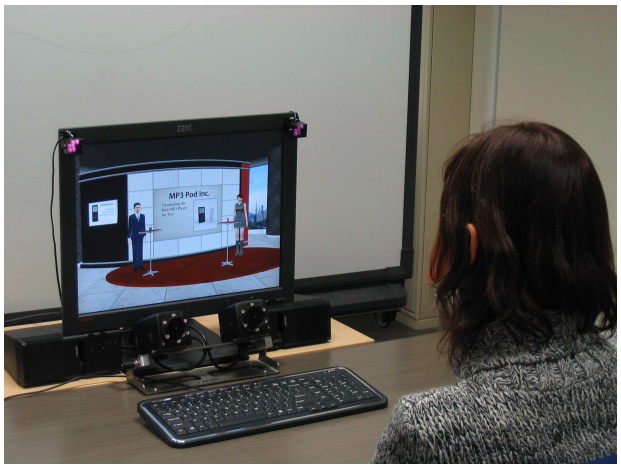

Figure 1: System setup.

In addition to interest estimation, we also implemented an automatic preference detection algorithm, which is applied to decision situations such as"Which item do you prefer?". Specifically, 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 [13]. 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 real-time preference detection component (AutoSelect) was tested in a study where users had to select their preferred necktie from two presented neckties through eye gaze [2]. The AutoSelect system achieved an accuracy of 81%.

4. GAZE-CONTINGENT PRESENTATION

The presentation scenario consists of a virtual sales scenario where a team of two 3D animated agents presents MP3 players to a human user. A professional Japanese character designer for "digital idols" created two highly realistic and expressive 3D agents (female and male), based on the appearance of two famous Japanese actors. Each character can perform body and facial gestures (emotional expressions), speak with proper lip-synchronization and direct its gaze at any specified scene entity as well as the user seated in front of the computer display screen. Agents and environment are controlled by MPML3D [6], a reactive framework that supports anytime user interaction, including real-time interpreted input from the eye tracker.

4.1 Eye Tracking Setup

A user is seated in front of a 30 inch screen (distance 80 cm) and stereo cameras of the faceLAB eye tracker from Seeing Machines [11]. The cameras and speakers are located below the screen. Two infrared pods are attached at the upper part of the display for illumination of the eyes (see Fig. 1). Calibration has to be performed for each user. The

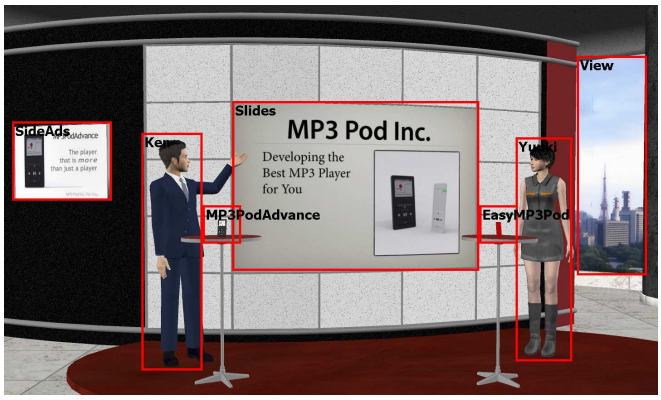

Figure 2: Screen areas of interest objects.

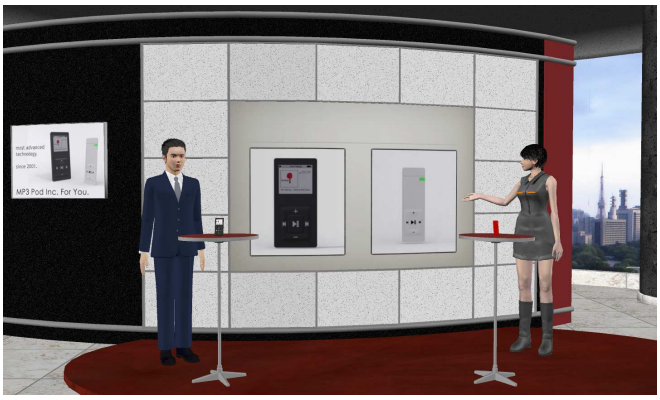

Figure 3: User is instructed to select an MP3 player.

resulting profile can be stored for reuse with the same user. The faceLAB software allows us to extract the coordinates of gaze points on the screen. For visualization purposes, the recorded data can be processed and analyzed with a screen-based analysis tool such as GazeTracker.

4.2 Interest Objects and Grounding

Each agent introduces an MP3 player by describing its features and advantages. The female agent ("Yuuki") promotes the EasyMP3Pod and the male agent ("Ken") promotes the MP3PodAdvance. During the presentation, the eye-based system monitors user interest in predefined screen objects. Specifically, the system analyzes whether the user attends to the dynamics of the presentation, which is based on alternately speaking agents and changing slides.

Screen areas that may trigger a system response when being looked at (or not looked at) are called 'interest objects'. Figure 2 shows the interest objects defined in our presentation setting. From left to 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 to the right. For each interest object, the IScore is calculated every frame (approx. 50/sec). When the score exceeds the threshold, the object becomes 'activated' and the agent(s) will react if a reaction is defined.

The key functionality of the presentation system is to monitor whether grounding is successful or not. 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' [3]. During the presentation, agents repeatedly apply indicative (deictic) gestures in order to establish referential identity. As fully embodied agents, they can perform pointing gestures to indicate the referent, such as the slide or one of the two virtual MP3 players. Grounding is considered successful if one of the following conditions is met: (1) the user's gaze shows a transition from the screen area of the speaking agent to the screen area of the referent during the performance of an utterance (or gesture) for at least 150 milliseconds, or within one second after the utterance (or gesture) terminated; (2) the user already attends to the referent. Since the agents also look at the referent when performing a deictic gesture, we might alternatively call successful grounding a state of "joint attention" [16] of user and agent.

When positive evidence in grounding is observed, the presentation will continue. In the case of negative evidence in grounding (i.e. the absence of positive evidence), the agents will interrupt their presentation and perform a 'alert' or 'suspension' response. In case of *alert* the co-presenter requests the user to focus on the current content of the presentation (the referent). *Suspension* means that the (current) co-presenter asks the (current) presenter to suspend the presentation and explains the object of the user's visual interest, such as the side advertisement, the view, or the co-presenter.

Currently, interruptions are handled in a simple way. In the case of alert, the presenter simply resumes the presentation after asking the user to pay attention to it. In the case of suspension, the presentation will be suspended at first by providing information about the respective interest object, and subsequently, the co-presenter agent will try to redirect the user to the presentation content.

4.3 Preference Formation

Towards the end of the presentation, the agents instruct the user to choose his or her preferred MP3 player. This scene is implemented by showing a slide as in Fig. 3. Each agent asks if the user prefers its presented MP3 player (EasyMP3Pod or MP3PodAdvance) and performs an accompanying deictic gesture to the player on the slide. A pre-study showed that the gaze cascade phenomenon will occur naturally in this situation. Users alternately look at the left part and the right part of the slide, and eventually exhibit a bias for a player in one part. Our system computes the decision within seven seconds (an empirical value taken from [13]). Here we do not intend to replace speech input (which appears more natural in this context) by gaze, but to investigate the possibilities of a gaze-based system. In the current version, the user is requested to press a key for indicating the choice. Then the agent with the preferred player expresses happiness about its successful promotion.

5. DISCUSSION

There is ample evidence that agent-based presentations are effective as an information medium [8] and also entertaining for the user [10]. Here we want to address the question how gaze input can improve the user's interaction experience. First it is noted that speech certainly conveys the richest information in human–computer interaction. However, speech might not be the preferred input modality for

scenarios such as presentations, which do not assume verbal expressions of interest (or non-interest) from the audience. Gaze, on the other hand, can be used to sense the user's interest and intention from involuntary eye movements. Moreover, the estimation of gaze is an unobtrusive and robust method to estimate user attention *continuously* and hence a gaze-based system can adapt its behavior anytime.

User awareness and attentiveness to the user's interest state [18, 12] are important to improve the user's interaction experience. Attentive interfaces provide the user with subtle control over system behavior without having to issue commands explicitly. Another term in use is *engagement* "...the process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in the interaction, evaluating whether to stay involved, and deciding when to end the connection." [14, p. 78] To some extent, this characterization of engagement applies to presentation situations as well. Here the audience agrees to 'collaborate' with the presenter by following the presentation.

6. CONCLUSIONS

We have described an eye-based infotainment presentation system featuring two highly realistic and expressive virtual 3D agents that are capable of responding to a user's focus and shift of attention and interest. The real-time analysis of eye gaze offers a powerful method to adapt the presentation to the user. The agents may alert the user if positive evidence for the successful grounding process is absent, e.g. when the user does not follow an indicative gesture of the agent, and estimate the user's preference for one screen object among two alternatives. For interest estimation, the system relies on a previously developed algorithm [9]. User preference estimation is realized by an automated version of the 'gaze cascade' effect, a finding from neuroscience [13]. How to handle situations where the system fails to estimate user (non-)interest and preference correctly remains an open issue and is left for future research.

We are currently in the process of conducting an extensive study to evaluate the system described in this paper. We hope to clarify the advantages of an eye-based system over a system that lacks this functionality. We are expecting results for several dimensions related to users' positive experience of the interface, such as involvement, engagement, and a feeling of co-presence.

Acknowledgements

The research was supported by a JSPS Encouragement of Young Scientists Grant (FY2005–FY2007) and an NII Joint Research Grant with the University of Tokyo (FY2006). The second author was supported by an International Internship Grant from NII under a Memorandum of Understanding with the University of Augsburg.

7. REFERENCES

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

[2] N. Bee, H. Prendinger, A. Nakasone, E. André, and M. Ishizuka. AutoSelect: What You Want Is What You Get. Real-time processing of visual attention and affect. In *Tutorial and Research Workshop on Perception and Interactive Technologies (PIT-06)*, pages 40–52. Springer LNCS 4021, 2006.

[3] H. H. Clark and S. E. Brennan. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. APA Books, WA, 1991.

[4] M. Maybury, O. Stock, and W. Wahlster. Intelligent interactive entertainment grand challenges. *IEEE Intelligent Systems*, 21(5):14–18, 2006.

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

[6] M. Nischt, H. Prendinger, E. André, and M. Ishizuka. MPML3D: a reactive framework for the Multimodal Presentation Markup Language. In *Proceedings 6th International Conference on Intelligent Virtual Agents (IVA-06)*, Springer LNAI 4133, pages 218–229, 2006.

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

[8] H. Prendinger, C. Ma, and M. Ishizuka. Eye movements as indices for the utility of life-like interface agents: A pilot study. *Interacting with Computers*, 19(2):281–292, 2007.

[9] P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of CHI'05)*, pages 221–230. ACM Press, 2005.

[10] T. Rist, E. André, S. Baldes, P. Gebhard, M. Klesen, M. Kipp, P. Rist, and M. Schmitt. A review of the development of embodied presentation agents and their appication fields. In Prendinger and Ishizuka [7], pages 377–404.

[11] Seeing Machines. Seeing Machines, 2005. URL: `http://www.seeingmachines.com/`.

[12] T. Selker. Visual attentive interfaces. *BT Technology Journal*, 22(4):146–150, 2004.

[13] S. Shimojo, C. Simion, E. Shimojo, and C. Scheier. Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322, 2003.

[14] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh. Where to look: A study in human–robot engagement. In *International Conference on Intelligent User Interfaces*, pages 78–84. ACM Press, 2004.

[15] I. Starker and R. A. Bolt. A gaze-responsive self-disclosing display. In *Proceedings of CHI'90*, pages 3–9. ACM Press, 1990.

[16] B. M. Velichkovsky and J. P. Hansen. New technological windows into mind: there is more in eyes and brains for human-computer interaction. In *Proceedings of CHI'96*, pages 496–503. ACM Press, 1996.

[17] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of CHI'01*, pages 301–308. ACM Press, 2001.

[18] S. Zhai. What's in the eyes for attentive input. *Communications of the ACM*, 46(3):34–39, 2003.