# Visual Interest Contingent Presentation Agents

Helmut Prendinger[1], Arjen Hoekstra[2],, Nikolaus Bee[3], Michael Nischt[3], and Mitsuru Ishizuka[4]

[1] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
helmut@nii.ac.jp
[2] Computer Science, Human Media Interaction
University of Twente, PO Box 217, 7500 AE Enschede The Netherlands
a.h.hoekstra@student.utwente.nl
[3] Institute of Computer Science, University of Augsburg
Eichleitnerstr. 30, D-86135 Augsburg, Germany
nikolaus.bee@gmail.com,Michael.Nischt@gmail.com
[4] 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.** This research proposes an interactive presentation system that employs eye gaze as an intuitive and unobtrusive input modality. Attentive behavior is an excellent clue to users' (visual) interest and preference. By analyzing and interpreting human eye movements in real-time, the interface can adapt to the current interest of the user and provide a more personalized or 'attentive' experience of the presentation to the user. Our system implements a virtual presentation room, where research content is presented by a team of two highly realistic 3D agents in a dynamic, interactive, and story-like way. The agents are capable of adapting their presentation according to the information from a non-contact video based eye tracker, thereby accounting for the user's focus and shift of visual interest in the presented material. A demo video based on this system won the GALA Award[5] as the best application of life-like agents at an international event.

## 1 Introduction

The challenge of giving a good presentation is to provide relevant and interesting content in an easily accessible way while keeping the attention of the audience during the entire presentation time. Human presenters often obtain feedback from the audience regarding their level of attention by simply looking at their behavior, specifically whether they are looking at the currently presented material, typically visualized on slides, at the presenter or somewhere else. If a presenter, e.g. a museum guide, observes that the attention of the spectators is diverted by other objects, he or she will try to adapt the presentation by taking the interest shift of the audience into account.

[5] http://hmi.ewi.utwente.nl/gala/

Since virtual presenter agents mostly do not have any input modalities available to determine the user's current focus of attention or interest, we propose a system that is based on human eye movements. As an input modality, eye gaze has the advantage of being an involuntary signal that reflects the user's visual interest [14]. Furthermore, the signal is robust and can be assessed accurately [4]. Speech, by contrast, certainly conveys the richest information in human–computer interaction, but it is not necessarily the preferred modality for scenarios such as presentation settings, which typically do not assume verbal expressions of interest from the audience.

Our proposed system can be conceived as reviving the "self-disclosing display" concept introduced in Starker and Bolt [19], where eye gaze is utilized as an input modality to recognize and respond to a user's interest. Their system would zoom in to areas of user interest and provide explanations via synthesized speech. Our work extends this concept by detecting both user interest and preference, and by embodied agents rather than a disembodied voice.

The remainder of this paper is structured as follows. Section 2 discusses related work concerning digital interactive storytelling, life-like characters, and eye gaze in human–computer interaction. In Section 3, the eye gaze based method to determine users' interest will be presented. Section 4 demonstrates our method by means of an implemented example. The paper is discussed and concluded in Sections 5 and 6, respectively.

## 2 Background and Related work

Rist et al. [15] describe the evolution of virtual presenters, starting from a single presenter not capable of handling user interaction to multiple presenters that can interact with multiple users. The most recent stage in the evolution described in [15] is very challenging and until now, it has not been realized in a satisfiable way. Our system takes one step back from this scenario, and features a presentation team (consisting of two agents) that is capable of adapting its presentation depending on the user's visual interest. al. [20] is similar to our system in that virtual agents can interrupt the presenter agent but unlike ours, they do not act as co-presenters. Furthermore, user interaction is not yet integrated into this project.

**Interactive Storytelling**.

With regard to the storytelling component, we do not have a dedicated engine as e.g. Cavazza et al. [3] or Iurgel [8]. Cavazza et al. [3] developed a character-based interactive storytelling system where the user can influence the story flow by giving advice to the characters or by 'physically' manipulating objects in the scene. Iurgel [8] introduced the mixed-reality art environment Art-E-fact in which virtual characters discuss art objects. The user can interact with the system by performing gestures in order to uncover hidden layers of paintings or to zoom in or out. Besides gesture recognition, the keyboard can be used as an input modality for chatting with the virtual agents. Story interaction is accomplished by a narration engine that handles scene based adaptation (exchanging, adding

or deleting parts of the story), improvisation based adaptation (adaptation of the actors' behavior), and levels of anticipation (e.g. distinguishing between 'normal' and 'abnormal' input).

Currently, our system does not provide a dedicated narrative component. While the presentation or story flow can be suspended in order to insert a sub-presentation related to the user's presumed interest object (derived from eye movements), our resumption strategies for continuing the presentation are straightforward, i.e. not selected by a narrative engine.

**Life-like Characters**. Life-like characters are virtual animated agents that are intended to provide the illusion of life or 'suspend disbelief' [1] such that users interacting with those agents will apply social interaction protocols and respond to them as they would to other humans, e.g. by listening to their story. Life-like characters have been shown to serve multiple purposes successfully; besides presenters, they can act as tutors, actors, personal communication partners, or information experts [13, 11].

In our system, two agents were designed based of the appearance of two famous Japanese actors. Their purpose is to act as presenters of research content accomplished at our institute, the National Institute of Informatics in Tokyo. In order to support their life-likeness, the agents are highly expressive. They can perform various gestures, such as greeting and counting, or 'beat', and deictic gestures. Besides body gestures, mimics for joy, surprise, and sadness are currently available. The Loquendo TTS (Text-To-Speech) engine is used to create natural sounding audio files [10]. Speech is combined with proper lip synchronization, and the head of the agents can be adjusted to any (physically natural) direction, e.g. to the direction of the other agent when giving turn or to the virtual slide. Nischt et al. [12] developed a scripting language called MPML3D that enables easy scripting of agent behavior within an reactive framework, based on the Multi-modal Presentation Markup Language (MPML) family [7].

**Eye Gaze**. Eyes are an intriguing part of the human face that are sometimes even seen as "windows to the soul". Heylen [6] investigated the major functions of eye gaze, including paying and signalling attention, conversation regulation, and demonstration of intimacy. When listening to a presentation is concerned, paying attention to its visualized content of key importance. The audience will also focus on the presenter's face to increase comprehension via perception of lip movements in addition to speech.

Recent attempts to integrate eye behavior into interactive systems are reported in Selker [17] who discusses the use of eye tracking in various applications, so-called 'visual attentive interfaces', such as the Magic Pointing and InVision systems. Magic Pointing lets the mouse pointer follow the user's eye gaze to speed up context changes on the screen. The InVision system analyzes eye patterns in order to prioritize activities. Here, the order in which people look at objects is used to determine what the user would like to do next. When looking at the a virtual kitchen the system could e.g. find out whether the user is hungry. The difference between those two applications is that Magic Pointing depends on user's conscious eye gaze, whereas InVision exploits unconscious gaze behavior.

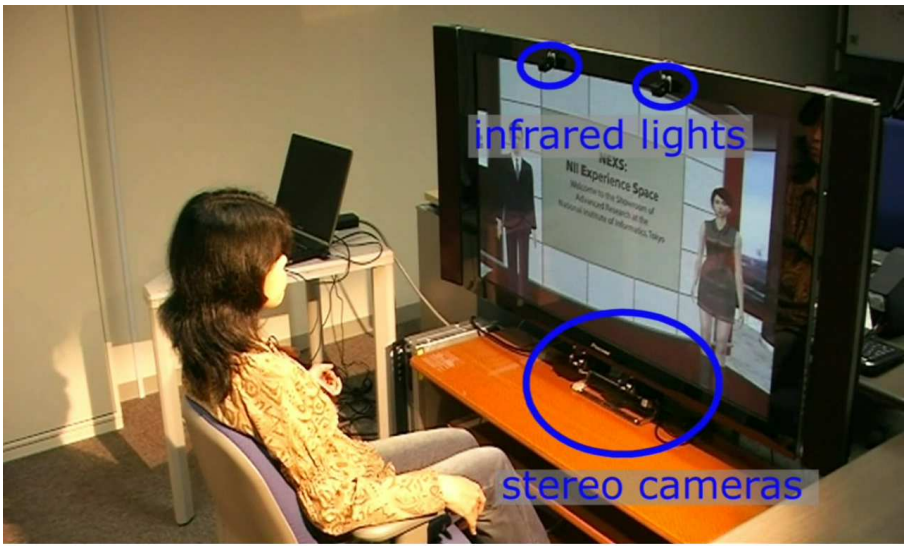


**Fig. 1.** Eye tracking setup.

The latter type of eye movement is also utilized in our presentation system, which is capable of recognizing a user's focus of attention and adapt the presentation accordingly in a non-command fashion, i.e without requiring the user to actively instruct the system.

## 3 Interest and Preference Estimation by Eye Tracking

In order to determine the users' visual interest in interface objects, eye gaze information is processed. The faceLab eye tracker [16] is used, which can track eye movements through real-time video footage from two cameras. Calibration for each user has to be performed only once. Our setup for the eye tracking system is depicted in Fig. 1. Here, the user is seated in front of a Pioneer PDP-505P 50 inch plasma screen with two stereo cameras positioned at the bottom of the screen. Two infrared pods are attached at the upper part of the display for illumination of the eyes. The eye tracking software allows us to extract the coordinates of gaze points on the screen.

### 3.1 Interest Estimation

The focus of interest is determined by a slightly modified version of the algorithm of Qvarfordt and Zhai [14]. These authors implemented an intelligent virtual tourist information environment (iTourist), for which they propose a new interest algorithm based on eye gaze. Two interest metrics were developed: (i) the Interest Score (IScore) and (ii) the Focus of Interest Score (FIScore).

The IScore refers to the object 'arousal' level, i.e. the likelihood that the user is interested in the (visual) object. When the IScore 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, computing the IScore is sufficient as we are mainly interested in whether the user's attention is currently on a certain object.

The basic component for the IScore is defined as follows [14]:

$$p = \frac{T_{ISon}}{T_{IS}}$$

$T_{ISon}$ refers to the accumulated gaze duration within a time window of size $T_{IS}$. In our application, $T_{IS} = 1000$ milliseconds was used. Hence, the longer some object is attended to, the higher the chance that the user would be interested in the object. This model of user interest is certainly simplistic as there can be other factors that enhance or inhibit interest. Therefore the following extended version of the previous formula was used [14].

$$p_{is} = p\,(1 + \alpha\,(1 - p))$$

In this formula, $p_{is}$ denotes the arousal level of an object, the so-called IScore ($0 \leq p_{is} \leq 1$). For our presentation system the threshold of 0.45 was chosen, which is an empirically derived value [14]. $p$ ($0 \leq p \leq 1$) is the eye gaze intensity (outcome of the previous formula) and $\alpha$ is the object's excitability modification factor ($-1 \leq \alpha \leq 1$). The modification factors are modelled as follows [14]:

$$\alpha = \frac{c_f\alpha_f + c_c\alpha_c + c_s\alpha_s + c_a\alpha_a}{c_f + c_c + c_s + c_a}$$

The terms in this formula are defined as:

- $\alpha_f$ is the frequency of the user's eye gaze 'entering' and 'leaving' the object ($0 \leq \alpha_f \leq 1$),
- $\alpha_c$ is the categorical relationship with the previous active object ($\alpha_c = -1|0|1$),
- $\alpha_s$ is the average size of all possible interest objects compared to the size of the currently computed object ($-1 \leq \alpha_s \leq 1$),
- $\alpha_a$ encodes whether the object was previously activated ($\alpha_a = -1|0$), and
- $c_0$, $c_f$, $c_c$, $c_s$, and $c_a$ represent empirically derived constant values of the corresponding factors. Some of these factors are domain dependent and are thus not applicable in all contexts.

The factors $\alpha_c$ and $\alpha_a$ were not (yet) integrated to our system. $\alpha_c$ concerns (semantic) relations between objects; $\alpha_a$ can be used to make the system respond in a different way when an object is activated multiple times.

We continue by explaining $\alpha_f$ and $\alpha_s$, the two remaining factors. $\alpha_f$ is represented as $\alpha_f = \frac{N_{sw}}{N_f}$, 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. When the user's gaze switches to some object many times, the value of the modification factor will increase and hence there will be a higher chance on excitation. $\alpha_s$ is represented by $\alpha_s = \frac{S_b - S}{S}$, whereby $S_b$ represents the average size of all objects and $S$ denotes the size of the currently computed object. This modification is intended to compensate for the differences between the size of the potential interest objects. Due to some noise in the eye movement signal, larger objects could have a higher chance of being 'hit' than smaller ones, which should be avoided.

### 3.2 Preference Estimation

Besides employing the IScore metric for detecting the user's interest, we also estimate the so-called 'gaze cascade' effect to determine the user's preference for situations involving a two-alternative forced choice. The gaze cascade effect was discovered in a study by Shimojo et al. [18], where users had to choose the more attractive face from two faces. They could demonstrate that there was a gaze bias towards the chosen stimulus in the last one and a half seconds before the decision was made. Bee et al. [2] developed a real-time system based on the gaze cascade effect that can estimate the user's preference with high accuracy. In our presentation system the gaze cascade effect was exploited to determine how the user wants the presentation to continue.

Details about the use of the gaze cascade effect and examples of the use of the interest algorithm will be given in the next section.

## 4 Examples of Handling User Interest and Preference

Our implemented system involves a team of two presentation agents that introduce the user to research of the National Institute of Informatics (NII) in Tokyo. Three topics were prepared: (i) a general introduction to NII and its mission, (ii) the research of Seiji Yamada on query expansion, and (iii) the research of Shin'ichi Satoh on face detection. Initially, the agents perform a self-introduction. During the presentation, the agents will adapt the presentation flow to a user in two ways, based on his or her eye gaze:

1. If the user shows interest in a particular interface object (an 'interest object'), the agents will interrupt their pre-defined presentation and respond to the user by providing information about the object the user shows interest in.
2. In order to determine the user's preference with respect to the next topic, the gaze cascade effect is computed at certain decision points, where the agents ask the user to select his or her preferred topic.

It is important to note that in both cases, non-conscious eye behavior of the user is analyzed, i.e. we do not instruct the user to perform a particular gaze behavior. While this is obvious in the first case, our previous experience [2] also demonstrated that users will naturally compare and select interface objects when confronted with a visual choice. Fig. 2 depicts the beginning portion of the choices of the predefined presentation flow.

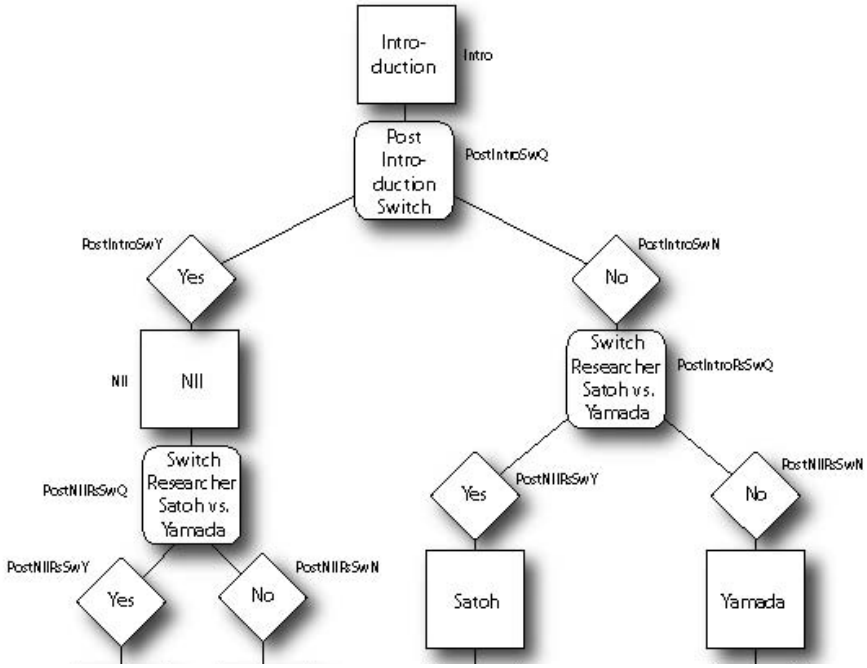


**Fig. 2.** Initial part of the presentation flow.

### 4.1 Responding to Interest

Currently, the following interest objects are defined in our scenario (see also Fig. 3): (a) the male agent 'Ken' (Boy), (b) the female agent 'Yuuki' (Girl), (c) the left-hand part of the slide, (d) the right-hand part of the slide, (d) the whole slide, (e) the logo of NII (National Institute of Informatics), and (f) the view out of the window to the right (from the user's perspective).

For each interest object the IScore is calculated every second. When the score exceeds the threshold, the object becomes 'activated' and the agent(s) will react (if a reaction is defined). Agent responses (or non-responses) are defined for three types of situations.[6]

[6] A demo video can be found at `http://research.nii.ac.jp/~prendinger/GALA2006/`


**Fig. 3.** The interest objects of our virtual environment.


(a) User is interested in slide content. (b) User is interested in view.

**Fig. 4.** Examples of user's estimated interest, indicated by (screen heat) gaze trails.

| | |
|---|---|
| Ken: | The transductive learning makes up for the smallness of user feedback. The transducer assigns labels from which the relevance is unknown with the same label as neighboring terms with known relevance. |
| User: | *[IScore exceeds threshold of the "View" object and gets activated.]* |
| Yuuki: | Ken, hold on a second... I couldn't help noticing that you are admiring the view we have from NII at the city. You can see our building is very close to the Imperial Palace. All the greenery belongs to the park of the palace. Well, so much about our neighborhood. Let's go back to our presentation, but please concentrate this time. |

**Fig. 5.** Sample agent dialogue when the user is interested in the outside view.

1. *Continuation of presentation*: The user attends to the currently explained (part of a) presentation slide. (A slide can be partitioned into a picture part and a text part, see Fig. 3.) Since the user's interest in the slide content is desired, the agent will continue with the presentation. Fig. 4(a) depicts a situation where the user attends to the explanation of the male agent by gazing at the slide content.
2. *Interruption of presentation*: If the user is detected to be interested in an interface object that is not currently explained, the system chooses between two agent responses.
   (a) Suspension: Fig. 4(b) shows a situation where the user looks out of the (virtual) window rather than attending to the presentation content. Here, the co-presenter agent Yuuki asks her colleague Ken to suspend the research presentation and continues with a description of the view (see Fig. 5). Thereafter, Ken continues with the presentation.
   (b) Redirecting user attention: The presenter agents do not suspend the presentation to comply with the user's interest but the co-presenter (Yuuki) alerts the user to focus on the presentation content.

The existing implementation of our presentation system handles interruptions in a simple way. If a user's interest object is not the currently explained object (typically a slide), the presentation will be suspended at first by providing information about that object, and subsequently, the co-presenter agent will try to redirect the user to the presentation content.

### 4.2 Responding to Preference

At predefined points during the presentation, the user has to choose the next presentation topic (see the 'switches' in Fig. 2). The agents introduce the given choice while a slide such as the one depicted in Fig. 6 is shown. As demonstrated in our previous study [2], the gaze cascade phenomenon will occur naturally in this situation. Users will alternately look at the left part and the right part of the slide, each consisting of a title and a picture, and eventually exhibit a bias for one part. The decision process occurs within seven seconds. Thereafter, the presentation continues with the chosen topic.

Observe that the screen areas in Fig. 6 on which the selection process is based are more confined than the two partitions in Fig. 3. The reason is that we wanted to avoid the uncertainty resulting from situations where the gaze of the user is at or close to the separation line between the two alternatives.

## 5 Discussion

While gaze-contingent interfaces are getting increasingly popular [4], it remains an open question how 'reactive' an interface that uses eye gaze as an input should be. The problem of distinguishing between eye movements that are just explorative and those that are meant as an input is known as the 'Midas Touch'


**Fig. 6.** Determining the user's preference by exploiting the gaze cascade effect.

problem: "Everywhere you look, another command is activated; you cannot look anywhere without issuing a command." (Jacob [9, p. 156]). Our presentation system avoids the Midas touch problem by (i) strictly confining the screen areas that could yield an agent response (the interest objects), and (ii) calculating user interest based on a well-established metric [14].

A related question concerns the manner in which an interface should be manipulated by gaze behavior. Hansen et al. [5] propose the following types of interactivity:

1. *Change in information.* Objects of user interest are explained in more detail.
2. *Change in point-of-view.* The camera position changes to the interest object of the user.
3. *Change in storyline.* The sequence of story events is dependent on where the user attends to.

Our current system supports all of those possibilities (to some extent). For instance, when the user is interested in the virtual outside view, the agent provides additional explanation of the view and the (virtual) camera shifts to show a full screen image of the view. Interest objects can also be conceived as 'hyper-links' to particular scenes of a story. Our gaze cascade based selection method can be seen as a (restricted) implementation to decide the progress of the presentation.

## 6 Conclusions

The use of eye gaze offers a powerful method to adapt a presentation to the current interest of a user, i.e. make the presentation contingent to user interest. Eye gaze as an input modality is particularly beneficial when verbal feedback is either not assumed or difficult to provide. Most presentations given by lecturers or museum guides are one-way communications that can nevertheless be adaptive to the audience if a presenter is sensitive to the behavior of the audience, such as their exhibition of visual interest or non-interest. Furthermore, while

the audience certainly has interest in specific presentation topics or parts of the presentation, it is unusual (or at least impolite) to verbally point out objects of interest repeatedly during the presentation. The online analysis of eye behavior thus provides an unobtrusive method to estimate user interest continuously.

In this paper, we have described a presentation system that features two virtual 3D presentation agents capable of responding to a user's focus of attention and interest in a natural way. The agent system [12] supports synchronized speech and lip movements, timed gestures, mimics, and head movements. In order to estimate user interest, the presentation system uses a previously developed algorithm [14], and has the presentation agents respond in an appropriate way. The gaze cascade effect [18, 2] is exploited at decision points in the presentation in order to determine with which presentation topic the user would like to continue.

In our future work, we will proceed along two research avenues. First, we plan to extend the interest estimation algorithm to cover relationships between interest objects in order to unveil e.g. a user's interest in comparing visual objects rather than choosing between them. Second, we intend to improve the presentation system by integrating narrative principles. This is important since currently, agent response to user input (visual interest) mostly 'interrupts' the presentation flow, which is thereafter simply resumed following the pre-defined storyline. It would be desirable to utilize user attention as a means to control the presentation in a natural and non-conscious way while preserving the narrative cohesion and persuasiveness of the presentation flow.

## Acknowledgements

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). The first author was supported by the JSPS Encouragement Grant. The third author was supported by an International Internship Grant from NII under a Memorandum of Understanding with the Faculty of Applied Informatics at the Univ. of Augsburg. We would also like to thank Dr. Ulrich Apel (NII) for scripting the dialogues.

## 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. M. Cavazza, F. Charles, and S. J. Mead. Interacting with virtual characters in interactive storytelling. In *Proceedings First Conference on Autonomous Agents and Multiagent Systems (AAMAS-02)*, pages 318–325, New York, 2002. ACM Press.

4. A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, London, UK, 2003.
5. J. P. Hansen, T. Engell-Nielson, and A. J. Glenstrup. Eye-gaze interaction: A new media – not just a fast mouse. In *The Second Swedish Symposium on Multimodal Communication*, 1998.
6. D. Heylen. A closer look at gaze. In *Proceedings AAMAS-05 Workshop on Creating Bonds with Embodied Conversational Agents*, pages 3–9, 2005.
7. M. Ishizuka and H. Prendinger. Describing and generating multimodal contents featuring affective lifelike agents with MPML. *New Generation Computing*, 24:97–128, 2006.
8. I. Iurgel. From another point of view: Art-E-Fact. In *Proceedings of Technologies for Interactive Digital Storytelling and Entertainment (TIDSE-04)*, pages 26–35, 2004.
9. R. J. K. Jacob. The use of eye movements in human-computer interaction techniques: What You Look At is What You Get. *ACM Transactions on Information Systems*, 9(3):152–169, 1991.
10. Loquendo Vocal Technology and Services, 2006. URL: `http://www.loquendo.com`.
11. A. Nijholt. Towards the automatic generation of virtual presentation agents. *Information Science Journal*, 9, 2006.
12. 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)*, 2006.
13. H. Prendinger and M. Ishizuka, editors. *Life-Like Characters. Tools, Affective Functions, and Applications*. Cognitive Technologies. Springer Verlag, Berlin Heidelberg, 2004.
14. P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI-05)*, pages 221–230. ACM Press, 2005.
15. 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 H. Prendinger and M. Ishizuka, editors, *Life-like Characters. Tools, Affective Functions and Applications*, Cognitive Technologies, pages 377–404. Springer, Berlin Heidelberg, 2004.
16. Seeing Machines. Seeing Machines, 2005. URL: `http://www.seeingmachines.com/`.
17. T. Selker. Visual attentive interfaces. *BT Technology Journal*, 22(4):146–150, 2004.
18. S. Shimojo, C. Simion, E. Shimojo, and C. Scheier. Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322, 2003.
19. I. Starker and R. A. Bolt. A gaze-responsive self-disclosing display. In *Proceedings CHI-90*, pages 3–9, 1990.
20. H. van Welbergen, A. Nijholt, D. Reidsma, and J. Zwiers. Presenting in virtual worlds: Towards an architecture for a 3D presenter explaining 2D-presented information. In *Proceedings First International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN-05)*, pages 203–212, 2005.