Dynamic Bayesian Network Based Interest Estimation for Visual Attentive Presentation Agents

Boris Brandherm
National Institute of Informatics
2–1–2 Hitotsubashi,
Chiyoda-ku,
Tokyo 101–8430, Japan
boris@nii.ac.jp

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

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

In this paper, we report on an interactive system and the results ofa formal user study that was carried out with the aim of comparing two approaches to estimating users' interest in a multimodal presentation based on their eye gaze. The scenario consists of a virtual showroom where two 3D agents present product items in an entertaining way, and adapt their performance according to users' (in)attentiveness. In order to infer users' attention and visual interest with regard to interface objects, our system analyzes eye movements in real-time. Interest detection algorithms used in previous research determine an object of interest based on the time that eye gaze dwells on that object. However, this kind of algorithm does not seem to be well suited for dynamic presentations where the goal is to assess the user's focus of attention with regard to a dynamically changing presentation. Here, the current context of the object of interest has to be considered, i. e., whether the visual object is part of (or contributes to) the current presentation content or not. Therefore, we propose to estimate the interest (or non-interest) of a user by means of dynamic Bayesian networks that may take into account the current context of the attention receiving object. In this way, the presentation agents can provide timely and appropriate response. The benefits of our approach will be demonstrated both theoretically and empirically.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H5.2 [**Information Interfaces and Presentation (e. g., HCI)**]: User Interfaces—*input devices and strategies, interaction styles, theory and methods*

General Terms

Human factors

Cite as: Dynamic Bayesian Network Based Interest Estimation for Visual Attentive Presentation Agents, Boris Brandherm, Helmut Prendinger and Mitsuru Ishizuka, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons (eds.), May, 12-16., 2008, Estoril, Portugal, pp. 191-198.

1. MOTIVATION

The recent progress in multi-modal interfaces facilitates new types of interactive applications, such as virtual games, audience-guided movies, and virtual travel guides [10, 7].

A promising direction is to process users' eye gaze as an indicator of their interest in the multi-modal application. This research direction was initially explored in 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. More recent works include the so-called 'Attentive User Interfaces' (AUIs) [17] or "visual attentive interfaces" [13], that aim to recognize the user's intention from natural gaze behavior. For instance, the InVision [13] 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.

Other works on gaze based interfaces focus on the regulation of conversational flow in a multi-agent environment. The FRED system [16] 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. If combined with speech, the agents can determine if they have to listen to someone else or if they can talk.

Our research is similar to the system described in [8], where the user converses with the virtual agent in the MACK system. Here, a head tracker is used to determine a user's gaze in a direction-giving task. The agent explains directions on a map and monitors the user's head. In that application, lack of negative feedback indicates successful grounding. 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.

In previous work [4] we analyzed and interpreted eye movements by means of a slightly simplified version of the algorithm introduced in [11] which has been developed for the virtual tourist information environment (iTourist). This algorithm determines the object of interest based on the eye gaze dwell time on that object. We experienced that this is not well suited for dynamic presentations where the goal is to assess the user's focus of attention with regard to a dynamically changing presentation. We realized that the current

context of the object of interest has to be considered, i. e., whether the visual object is part of (or contributes to) the current presentation content or not. Therefore, we propose to estimate the interest (or non-interest) of a user by means of dynamic Bayesian networks that may take into account the current context of the attention receiving object. In this way, the presentation agents can provide timely and appropriate response compared to the previous solutions.

The paper is structured as follows: we first describe our gaze-based application scenario (Sect. 2), where two virtual agents promote a fictitious product to the user. The main part of the paper consists in the comparison of the interest estimation algorithms, the IScore/FIScore metric [11] and our own metric, which is based on dynamic Bayesian networks (Sect. 3). Both metrics are first explained, and then compared with respect to their behavior and impact on spectators of the multi-modal presentation. For that purpose, a user study was conducted, which is described in Sect. 4. Some conclusions are drawn in Sect. 5.

2. EYE GAZE BASED PRESENTATION

The presentation scenario involves two three-dimensional (3D) animated agents in the role of virtual presenters of MP3 players.[1] Both agents can perform body and facial gestures (emotional expressions) and lip-synchronization. They can direct their gaze at any specified scene entity as well as the user seated in front of the computer display screen. MPML3D is used as a control language for the agents and the environment [9]. We are using the video based eye tracker faceLAB from Seeing Machines [12] to recognize users' gaze.

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. 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'. Fig. 1 shows the interest objects defined in our presentation setting. From left to right:

- 'SideAds', a total of four slides that advertise the MP3 players and are exchanged every five seconds;
- male agent 'Ken';
- 3D model of MP3PodAdvance;
- virtual slide;
- 3D model of EasyMP3Pod;
- female agent 'Yuuki';
- the view out of the window to the right.

For each interest object, the interest score is calculated every frame (60 times/sec). When the score for an object exceeds a threshold, the agent(s) will react if a reaction is defined.

[1]The presentation scenario is based on the one previously described in [4].

Figure 1: Bounding boxes of interest objects [4]. (The displayed computer screen area is clipped for convenience and does not show the user's actual view. The true view of the user is shown in Fig. 4.)

The presentation system monitors 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. The agents perform pointing gestures to indicate the referent, such as the slide or one of the two virtual MP3 players. 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 an '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.

Besides failed grounding situations, we also want to perform an interruption if the user attends to interest objects that are not considered as part of the current presentation content. Therefore, we have to take into account not only new gaze points but also the current context of the object. The 'context of an object' is determined by the content of the presentation and indicates whether some visual object belongs to the presentation content or not. If an agent talks about the content of a slide for an extended period of time, the user may look either at the slide or the agent. In this case, we say the user *follows* the presentation. In the latter case, an object is said to *distract* the user from the presentation.

In order to accommodate for these situations, we propose a new approach that estimates the interest (or non-interest) of a user by means of dynamic Bayesian networks.

3. INTEREST ESTIMATION

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. In the following, we will first present two interest estimation algorithms, and then compare their behavior regarding our application scenario.

3.1 IScore and FIScore

Qvarfordt and Zhai [11] developed two interest metrics for the virtual tourist information environment (iTourist): the Interest Score (IScore), and (2) the Focus of Interest Score (FIScore).

3.1.1 Characteristics

IScore denotes the likelihood that a user is interested in some visual object. When the IScore metric passes a certain threshold, the object is said to become 'active'. 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 (as the user lost interest in that object) and a new active object is selected based on its IScore.

3.1.2 Method

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} (e. g. 1000 ms). In order to account for factors that may enhance or inhibit interest, [11] 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. Our simplified version has two out of the four parameters defined in [11]: (i) α_f represents the frequency of the user's eye gaze 'entering' and 'leaving' an object, which indicates interest in that object; (ii) α_s represents the average size of all possible interest objects compared to the size of the currently computed object, which is intended to compensate for differences in the size of potential interest objects, and the related difference of being 'hit' by chance.

As in IScore, the basic component in FIScore is the gaze intensity on the active object. In addition, FIScore considers gaze intensity on other interest objects during a pre-specified time window. The time window for FIScore is larger than the one for IScore, e. g., twice as long.

3.2 NIIScore

The core algorithm of the NIIScore is based on dynamic Bayesian networks (DBNs) [1]. A DBN is an extension of a Bayesian networks (BNs), which is a widely used method for reasoning about uncertain knowledge [6]. The extension in DBNs refers to the possibility of modeling dynamic processes. Each time the DBN receives new evidence, a new time slice is added to the existing DBN. In principle, DBNs can be evaluated with the same inference procedures as (static) BNs, but their dynamic nature places heavy demands on computation time and memory. As more and more time slices are added to the DBN, the more and more computational resources (like time and memory) are necessary to solve the DBN. Therefore, roll-up procedures have to be applied that cut off old time slices without eliminating their influence on the newer time slices. In our system, roll-up is non-approximative and information preserving.

We employ the JavaDBN tool [1], which compiles a given DBN into source code as the basis for all computations, including inference, roll-up, and parameterization. In this way, the developer (knowledge engineer) does not have to care about computational complexity issues and can instead concentrate on the task of modeling the DBN. Since our system requires real-time operation, the source code is used, which circumvents the issue of garbage collection. Hence, unforeseeable interruptions can be avoided.

3.2.1 Characteristics

Before presenting the details of our DBN, we first discuss the requirements that should be satisfied by the new interest estimation method, which provide a strong case for the suitability of using DBNs.

In order to respond appropriately to the user, we need the following measures: (1) a measure for the interest of the user in the presentation; (2)a measure for the interest of the user in some particular screen object; and (3) measure of the time the user's eye gaze dwells on an object. In other words, we have to take into account not only the new gaze point but also the current context of the object and the preceding estimations of the object itself.

3.2.2 Method

In our system, each object of interest has an associated DBN.[2] With each new measurement of the eye tracker:

1. A new time slice is attached to the DBN of each interest object in the scene;
2. The inference is computed with regard to the new evidence – the user attend or does not attend to the object – and the contextual role of the object within the presentation;
3. A roll-up is performed for the previous time slice.

The context value of an object changes over time during the presentation and determines whether the user is supposed to attend to the object (in order to be able to follow the presentation properly) or not.

Below, we first show the structure of our DBN (the qualitative part) and then explain the conditional probabilities (the quantitative part) and thresholds used in the DBN.

The following are the core nodes of the DBN (see Fig. 2):

- 'Follows Presentation' (FP) represents the user's (general) interest in the presentation;
- 'Interest in object' (IIO) represents the user's interest in some particular object;
- 'Looks at' (LA) denotes the time that the user attends to some object.

As shown in Fig. 2, the Node LA is influenced by the previous value (score) and the current gaze point – independent of the context. On the other hand, the Nodes FP and IIO are influenced by the predecessor values, the current gaze point, and the context value as follows:

- If both "Context: User is supposed to look at object" and "Sensor gaze: User looks at object" hold, then the value of the Node FP should increase, whereas the value of the Node IIO should remain unchanged.
- If "Context: User is supposed to look at object" and "Sensor gaze: User does not look at object" hold, then the value of the Node FP should be decreased and the value of the Node IIO should remain the same.

[2]This idea was previously introduced in [2].

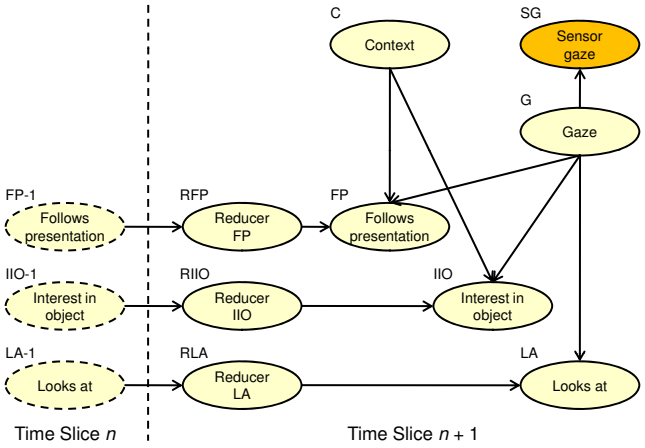

Figure 2: Time Slice $n+1$ with parent nodes of the preceding Time Slice n of the dynamic Bayesian network for the estimation of user's interest.

- On the other hand, if "Context: User is supposed not to look at object" and "Sensor gaze: User looks at object" hold, then the value of the Node FP should be decreased whereas the value of the node IIO should increase.

- If "Context: User is supposed not to look at object" and "Sensor gaze: User does not look at object", then the values of both Nodes FP and IIO should remain the same.

The Reducer nodes RFP, RIIO, and RLA model the decay from time slice to time slice. However, a given new observation can override the decay value ("priority of recency"). The Node 'Sensor gaze' models the reliability of the eye tracker. For simplicity, we assume 100% reliability.

In the following, we will demonstrate how to determine the conditional probabilities of the nodes in the transition model for our application (the presentation).

In our setup, the eye tracker performs 60 measurements per second, i. e., one time slice of the DBN corresponds to 16.67 ms. In order to encode the visual information of an object, the user has to focus at that object for at least 160 ms. Hence the score has to pass the threshold within eleven time slices ($(11-1) \times 16.67\,\text{ms} = 166.7\,\text{ms}$). In order to determine the values, the conditional probabilities of the variables 'Reducer FP' and 'Follows Presentation' were parameterized as follows:

$$\text{cpt}(\text{R\_FP}) = \left[\begin{array}{c|cc} \text{FP-1} & p_1 & p_2 \\ \hline r_1 & a_1 & a_2 \\ r_2 & a_2 & a_1 \end{array}\right],$$

$$\text{cpt}(\text{FP}) = \left[\begin{array}{c|cc|cc|cc|cc} \text{Gaze} & \multicolumn{4}{c|}{g_1} & \multicolumn{4}{c}{g_2} \\ \text{Context} & \multicolumn{2}{c|}{c_1} & \multicolumn{2}{c|}{c_2} & \multicolumn{2}{c|}{c_1} & \multicolumn{2}{c}{c_2} \\ \text{R\_FP} & r_1 & r_2 & r_1 & r_2 & r_1 & r_2 & r_1 & r_2 \\ \hline p_1 & 1 & 0 & 1 & b_2 & b_1 & 0 & 1 & 0 \\ p_2 & 0 & 1 & 0 & b_1 & b_2 & 1 & 0 & 1 \end{array}\right],$$

with $a_1 = a$ and $a_2 = 1-a$, $b_1 = b$ and $b_2 = 1-b$.

If the belief value of the preceding node FP in Time Slice n is

$$\text{Bel}(\text{FP}) = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix},$$

and the current Time Slice $n+1$ is updated with the evidence "Sensor gaze: User does not look at object" and "Context: User is supposed to look at object", the inference yields the updated belief value

$$\text{Bel}(\text{FP}) = \alpha \begin{pmatrix} 2-b \\ b \end{pmatrix}$$

for node FP. Here, α is a normalizing constant in order to guarantee that all values in the vector add up to 1. By repeating the process, we obtain

$$\text{Bel}(\text{FP}) = \alpha \begin{pmatrix} 2-2\,b+2\,a\,b+b^2-2\,a\,b^2 \\ 2\,b-2\,a\,b-b^2+2\,a\,b^2 \end{pmatrix}.$$

In order to systematically test the parameters a and b, we employed JavaDBN [1] to compute the belief value for the Node FP (evidence as above) as follows:

$$\text{Bel}(\text{FP}) = \alpha \begin{pmatrix} a\,e_1 + (1-a)\,e_2 + (1-b)\,((1-a)\,e_1 + a\,e_2) \\ b\,((1-a)\,e_1 + a\,e_2) \end{pmatrix},$$

whereby the believe value of the node FP in the preceding time slice is assumed as $\text{Bel}(\text{FP}) = \alpha \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$. Using this method, the belief value of the Node FP can be obtained iteratively for multiple time slices and different values of a and b. The values show the expected behavior for $a = 0.90$ and $b = 0.90$, with 0.64 as threshold.

In this experiment we just exploit the belief value of the Node IIO in the dynamic Bayesian network which takes into account the context of the associated interest object.

3.3 Comparison of the Interest Scores

In this section, we compare the behavior of both interest detection algorithms based on an example of a typical conversation in a dynamical changing presentation, where two virtual agents interact with each other. For demonstration purposes, the eye gaze behavior of the (imagined) user has been idealized.

We assume that a counter for agent Ken is defined in the presentation which increases by one (unit) if the value of the considered interest detection algorithm exceeds a certain threshold. The system is programmed to trigger a response after three times, e. g., Ken says "Is there something wrong with my necktie? You are looking at me even when Yuuki is talking." Obviously, this action should be triggered only if two conditions are met: (i) Yuuki is talking, and (ii) the user is looking at Ken.

The system behavior (specifically the values of the scores and the counter for Ken) is visualized in Fig. 3. The Arabic numerals correspond to the enumeration in the listing and the letters 'a' and 'b' correspond to the IScore and the NIIScore version, respectively.

The exemplary system behavior is as follows:

1. Yuuki is speaking and the user directs his/her attention to Yuuki.

   (a) Iscore for Ken is under the threshold. The counter for Ken has the value 0.

   (b) NIIScore for Ken is under the threshold. His counter has the value 0.

2. Yuuki is still speaking but the user directs his attention to Ken.

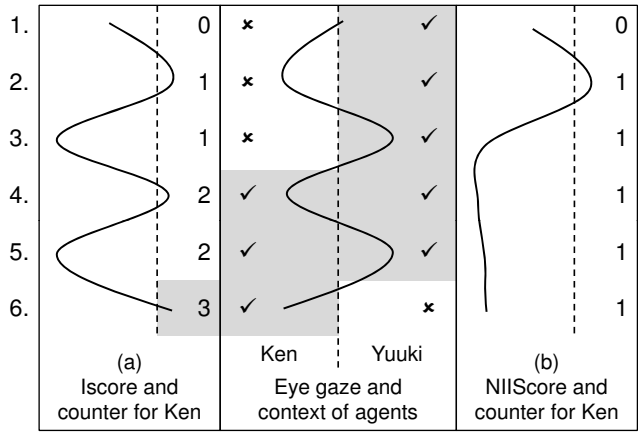

Figure 3: System behavior. The values of (a) IScore and (b) NIIScore for Ken, the corresponding counters, the eye gaze behavior of the user and the context of the agents Ken and Yuuki during the flow of the presentation are shown.

(a) Iscore for Ken increases and exceeds the threshold. The counter for Ken increases by 1 and has now the value 1.

(b) NIIScore for Ken increases and exceeds the threshold. The counter for Ken increases and has now the value 1.

3. Yuuki is still speaking and regains user's attention.

   (a) Iscore for Ken falls under the threshold.

   (b) NIIScore for Ken decreases and falls under the threshold.

4. Yuuki introduced Ken and the user is looking at him.

   (a) Iscore for Ken increases and exceeds the threshold. The counter for Ken increases by 1 and has now the value 2.

   (b) Ken is part of the current flow of the presentation such that the NIIScore for Ken doesn't increase. The counter for Ken remains at the value 1.

5. Yuuki is still talking about Ken. The user is looking at Yuuki.

   (a) Iscore for Ken falls under the threshold.

   (b) NIIScore for Ken doesn't change.

6. Yuuki hands over to Ken and the user looks at him.

   (a) Iscore for Ken increases and exceeds the threshold. The counter for Ken increases by 1. The counter for Ken has now the value 3. The action 'Ken says "Is there something wrong with my necktie? [...]"' is triggered by the system.

   (b) Ken is now part of the current flow of the presentation such that the NIIScore for Ken doesn't increase and therefore doesn't exceed a threshold. The counter for Ken stays at the same value 1.

In the example, we can easily see that the value of the NIIScore does not exceed the threshold as often as the value of the IScore. Furthermore, the IScore triggered an action in an inappropriate situation. The IScore is thus not well suited for interactive systems where we aim to determine a user's interest in a dynamically changing presentation. In this setting, the current context or role of the visual object within the presentation has to be considered, e. g., whether the object distracts the user from the presentation or not.

4. EXPERIMENTAL STUDY

4.1 Theory

The general hypothesis of our research on attentive presentation agents is that the recently introduced NIIScore can provide a more natural interaction experience as compared to the approach that is based on the IScore and FIScore metrics. Our hypothesis can be stated as follows:

> *If the interest detection algorithm considers the current context or role of the visual object within the presentation, users will experience the presentation agents as more mindful and exhibit more natural gaze behavior.*

We assume that users experience the interaction with attentive agents in a similar way as they do in human face-to-face communication, i. e., as more engaging, and inducing a sense of involvement and co-presence with the presenters. Our hypothesis is anchored in questions (from a questionnaire) regarding concepts of "face-to-face"(communication), "involvement", "co-presence", and "partner evaluation" proposed in [5], and "engagement" as described in [14].

4.2 Method

In this section we describe the experimental design, participating subjects, and the way the experiment has been conducted. The apparatus and the procedure are nearly the same as in our previous experimental study [4].

4.2.1 Experimental Design

The experiment had a between-subjects design in which the following two versions were implemented.

- Version based on IScore and FIScore (IScore Version): The system determines the user's interest by means of the IScore and the FIScore.
- Version based on dynamic Bayesian networks (NIIScore Version): The NIIScore is employed for interest estimation.

Note that all agent responses that may occur in the IScore version, can also occur in the NIIScore version (and vice versa). This allows us to compare the two versions, because there are no reactions which can only occur in one version.

4.2.2 Subjects

The study had nineteen subjects participating. They were all students, researchers, or staff from our institute, and received an amount corresponding to USD 9 for their attendance. There were technical difficulties with one subject, because the system crashed during run-time for unforeseeable reasons. Two subjects could not be calibrated because

of reflections of glasses and/or eye gaze aberrations. All of those subjects were excluded from the study beforehand. The age of the remaining sixteen subjects (six female, ten male) ranged from 22 to 59 yrs (average 31.8 yrs). They were randomly assigned to the IScore and NIIScore versions.

4.2.3 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 processor. The eye-tracking software faceLAB from Seeing Machines [12] 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 to the left and right of 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 (see Fig. 4).

Figure 4: System setup with stereo cameras of the eye tracker in the front, and two infrared pods attached at top of the screen.

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 and all eye gaze data has been logged in a separate data file. The schematic setup is shown in Fig. 5.

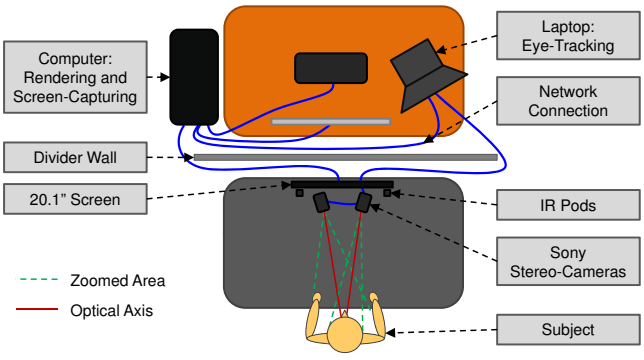

Figure 5: Experimental system setup. (Slight modification from [4].)

4.2.4 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 and to attend to the behavior of the agents. 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 tracking. For the calibration the experimenter has to follow the menu-based instructions of the faceLAB software. It is a step-by-step process where 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 four to six minutes. During that time, only the experimenter and an assistant were present in the room and silence was kept. After the presentation had been completed, the subjects filled in a questionnaire with nineteen questions that addressed their impression of the presentation. They were also briefly interviewed in an informal way.

4.3 Results

We first present some general results. The mean length of the presentation in the IScore and the NIIScore versions was 285 sec and 276 sec, respectively. In the IScore version the shortest presentation took 254 sec and the minimum length in the NIIScore version was 225 sec. The longest presentation was 333 sec in IScore version, and 340 sec in the NIIScore version. There were no significant differences between both versions concerning the length of the presentations.

4.3.1 Results for the Experience Dimensions

In order to test the 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-presence and agent evaluation have been borrowed from [5], the engagement dimension was derived from the description in [14]. The results of the questionnaires are listed in Table 1 ordered by the aforementioned dimensions.

Interestingly, most of the statistically significant results (four out of five) relate to questions in the co-presence and the engagement dimensions. None of the questions of the face-to-face and the involvement dimension showed significant results, and sometimes even showed the same mean value in both scores.

In both versions, the subjects did not feel as if they were in a real showroom for MP3 players. However, in the NIIScore version, there is a tendency that the subjects felt being more addressed as potential costumers of MP3 players than in the IScore version ($t(14) = -1.27$; $p = 0.11$).

Let us now have a closer look on the significant results. Concerning the co-presence dimension the subjects in the NIIScore version felt that the agents were aware of them to a significantly higher extent ($t(14) = -2.13$; $p < 0.05$) and subjects furthermore found that the agents paid closer attention to them to a significant higher extent ($t(14) = -2.33$; $p < 0.05$) than the subjects in the IScore version.

Table 1: t-test results (one-tailed) for face-to-face, involvement, co-presence, agent evaluation, and engagement.

	IScore		NIIScore			
Questions	*Mean*	*SD*	*Mean*	*SD*	$t(14)$	p
Face-to-face						
I could readily tell when the agents talked to me.	2.13	0.70	2.13	0.70	0.00	0.50
The conversation between the two agents seemed very natural.	1.00	2.00	1.25	2.21	-0.35	0.37
I felt that the two agents are a good team and communicate with each other well.	1.50	0.86	1.50	4.00	0.00	0.50
Sometimes I thought the agents react to me in a strange way.	1.25	2.50	0.88	3.56	0.43	0.34
The agents interrupted each other frequently and seemingly for no reason.	0.13	4.70	0.00	3.43	0.12	0.45
Involvement						
I really enjoyed listening to the presentation.	0.88	2.70	0.88	3.84	0.00	0.50
Co-presence						
I felt as if I were in a real showroom for MP3 players.	0.25	3.07	0.13	3.55	0.14	0.45
I felt that the agents were aware of me.	1.38	2.84	2.75	0.50	-2.13	< 0.05
I found that the agents paid close attention to me.	1.50	0.86	2.38	0.27	-2.33	< 0.05
Agent evaluation						
The female agent (Yuuki) was friendly.	1.00	3.14	0.63	2.84	0.43	0.34
The male agent (Ken) was friendly.	0.50	2.00	0.25	1.64	0.37	0.36
The female agent (Yuuki) did NOT take a personal interest in me.	-0.13	1.84	-0.75	3.64	0.76	0.23
The male agent (Ken) did NOT take a personal interest in me.	-0.88	2.13	-0.13	3.84	-0.87	0.20
I trusted the female agent (Yuuki).	0.63	2.84	2.00	1.71	-1.82	< 0.05
I trusted the male agent (Ken).	1.25	0.50	0.88	1.55	0.74	0.24
Engagement						
I felt that it was important to the agents that I am listening to them.	1.75	0.50	2.00	1.71	-0.48	0.32
I had the impression that the agents cared about my interest.	1.13	2.41	2.50	0.57	-2.25	< 0.05
I felt that the agents were aware that I am attentive to their presentation.	1.00	2.00	2.38	1.13	-2.20	< 0.05
I had a real sense of being addressed as a potential costumer of MP3 players.	0.75	2.50	1.75	2.5	-1.27	0.11

In the engagement dimension the subjects in the NIIScore version felt to a significantly higher extent ($t(14) = -2.20$; $p < 0.05$) that the agents were aware that they are attentive to their presentation. Additionally, they had the impression that the agents cared more about their interest in the NIIScore version than in the IScore version ($t(14) = -2.25$; $p < 0.05$).

The subjects in the NIIScore version trusted the female agent "Yuuki" to a significantly higher extent than in the IScore version ($t(14) = -1.82$; $p < 0.05$), but there was no significant result for the male agent "Ken" ($t(14) = 0.74$; $p = 0.24$). At the time of writing, we have no plausible explanation for this result.

Surprisingly, the subjects in the IScore version did not indicate that the agents acted in a strange way. In seven out of eight presentations, IScore triggered an inappropriate behavior: The user was looking at one agent because that agent was explaining something, but then the agent complained that the user is looking at him/her. (Such a behavior did not occur in the DBN version.) However, 5 of 7 persons pointed out the inappropriate agents responses when filling in the comment section of the questionnaire. Some of their remarks will be described below.

4.3.2 Informal Subject Comments

Two subjects (IScore and NIIScore version) suggested that the agents should smile at users for three different purposes: (i) smiling when being looked at would make the agents appear somewhat more polite, (ii) as a feedback channel for the users to let them know whether their gaze is recognized, and (iii) to attract the user in the role of a customer, e. g., by over-friendly behavior. These comments concern a design decision of our current implementation. If the user does not follow the presentation as expected, the agents will interfere by alerting the user. A future implementation might consider to provide positive feedback for expected user behavior, instead of negative feedback to the absence of expected behavior.

As mentioned previously, the subjects in the IScore version did not think to a significantly higher extent that the agents sometimes responded strange to them. Five out of the seven candidates who experienced this strange behavior (71%) mentioned it in the comments section of the questionnaire. Remarks included "Strange reaction of the agents: I had to look at them and then they ask 'Why do you look at me?' " and "Seemed strange that the agents reacted 'negatively' while I was looking at them and they were talking."

Two subjects in the NIIScore version complained that the agents were too sensitive: "The agents seemed to tell me off (chastise me) every time I glanced away for a split-second." and "I feel not free if the agents pay too much attention, if they are too sensitive. In a real presentation (human beings), I feel free to look out through the window for five seconds without interrupting the people." Here, the agents' behavior could be adjusted, e. g., by changing some values in the NIIScore such that over-sensitive agent reactions do not occur.

We conclude this section with the statement of a subject from the NIIScore version, who expressed excitement about the system with the words "It's very impressive. The agents can follow my eyes."

5. CONCLUSION

We presented an interactive system and the results of a formal user study that was carried out with the aim of comparing two approaches to estimating users' interest from eye gaze. The scenario consists of a virtual showroom where two 3D agents present product items in an entertaining way, and adapt their performance according to users' attentiveness, of lack of attentiveness.

The user study shows that our DBN-based interest recognition algorithm performs significantly better than an algorithm based on IScore and FIScore, specifically for the 'co-presence' and 'engagement' dimensions. The result can be explained by the fact that our interest detection algorithm considers the current context of the object of interest, i. e., whether the visual object is part of (or contributes to) the current presentation content or not, and thus contributes to more natural agent responses.

Certainly, the context parameter can also be built into other methods, such as IScore. Nevertheless, we believe that dynamic Bayesian networks provide a more unified and flexible solution, especially if we plan to consider additional parameters, either deriving from the presentation scenario, or from the user.

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), an NII Joint Research Grant with the Univ. of Tokyo (FY2006–FY2007).

6. REFERENCES

[1] B. Brandherm and A. Jameson. An extension of the differential approach for Bayesian network inference to dynamic Bayesian networks. *International Journal of Intelligent Systems*, 9(8):727–748, 2004.

[2] B. Brandherm, H. Prendinger, and M. Ishizuka. Interest estimation based on dynamic Bayesian networks for visual attentive presentation agents. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI-07)*, pages 346–349. ACM Press, 2007.

[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, Washington, 1991.

[4] T. Eichner, H. Prendinger, E. André, and M. Ishizuka. Attentive presentation agents. In *Proceedings 7th International Conference on Intelligent Virtual Agents (IVA-07)*, pages 283–295. Springer LNCS 4722, 2007.

[5] M. Garau, M. Slater, S. Bee, and M. A. Sasse. The impact of eye gaze on communication using humanoid avatars. In *Proceedings SIGCHI Conference on Human Factors in Computing Systems (CHI-01)*, pages 309–316. ACM Press, 2001.

[6] F. Jensen. *Bayesian Networks and Decision Graphs*. Springer, Berlin New York, 2001.

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

[8] 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.

[9] 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.

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

[11] 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.

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

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

[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 CHI-90*, pages 3–9. ACM Press, 1990.

[16] 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.

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