Evaluating the Interaction with Synthetic Agents Using Attention and Affect Tracking

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

Chunling Ma, Jin Yingzi,
Kushida Kazutaka, Mitsuru Ishizuka
Dept. of Information and Communication Eng.
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
ishizuka@miv.t.u-tokyo.ac.jp

ABSTRACT

We motivate an approach to evaluating the utility of synthetic agents that is based on human physiology rather than questionnaires. The primary tool is an eye tracker that provides quantitative evidence of a user's focus of attention. The secondary tool is a signal encoder for skin conductance and heart rate in order to gain insight into the user's affective state. The salient feature of our evaluation strategy is that it allows us to measure important properties of the user's interaction experience on a moment-by-moment basis. We describe an empirical study in which we compare attending behavior and affect of participants watching the presentation of an apartment by three types of media: a synthetic agent, a text box, and speech only.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Human Factors

Keywords

User study, eye tracking, synthetic agents, presentation

1. INTRODUCTION

While significant progress has been made in some aspects of synthetic agents, e.g. their visual appearance, evidence of their positive impact on human–computer interaction is still rare. A common feature of most evaluations of interface agents is that they are based on questionnaires and focus on users' experience with the systems hosting those agents, including questions about their believability, engagingness, or utility [1]. However, subtle aspects of the interaction, such as whether users pay attention to the agent or not, cannot be deduced reliably from self-reports. In this paper, we propose a different approach to evaluating synthetic interface agents. We will analyze the eye movements of users in order to obtain quantitative evidence of a their focus of attention. Specifically, we will track and analyze eye movements while users are following the online presentation of an apartment in three conditions (agent, text box, voice only). We will describe both spatial and temporal analyses of users' eye movements during the presentation.

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

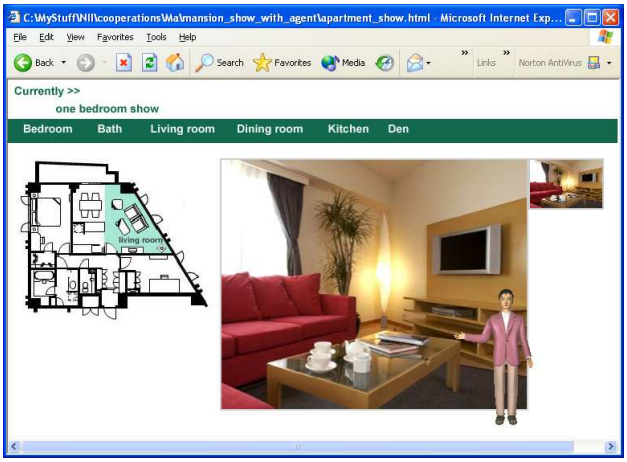

Figure 1: A synthetic agent presents the living room of the apartment.

2. METHOD

Experimental Design

A presentation of an apartment located in Tokyo has been prepared using a web page based interface. Views of each room of the apartment are shown during the presentation, including pictures of some parts of the room and close-up pictures. Three versions of the apartment show have been designed for the experiment: (i) *Agent (& speech) version.* A character called "Kosaku" presents the apartment using synthetic speech and deictic facial and hand gestures (see Fig. 1). The character is controlled by a version of MPML [3]; (ii) *Text (& speech) version.* The presentation content of each scene is displayed by a text box and read out by Mi-

crosoft Reader; (iii) *Voice (only) version.* Synthetic speech is the only medium used to comment on the apartment.

The main purpose of programming the Text and Voice versions was to provide interfaces that can be compared to the Agent version in terms of users' eye movements. The same type and speed of (synthetic) voice was used in all versions. Fifteen subjects, students and staff from the University of Tokyo, participated the study (5 in each version).

Procedure

The subjects were briefed about the experiment and instructed to watch the demonstration carefully. The subjects were first connected to the bio-sensors (SC, heart rate) and put on the cap with the eye tracker (NAC EMR-8B), and then calibration was performed for the eye tracker. After that, the subjects were shown the presentation, which lasted for 8 minutes. Finally, the subjects were freed from the eye tracking equipment, and asked to fill out a questionnaire concerning the presented material in order to report on their perception of the interface and the presentation (the questionnaire results are not described here).

Data Analysis

For analysis, the recorded video data of a presentation were first divided into individual scenes. A scene is a presentation unit where a referring entity (agent, text box, or voice) describes a reference object (an item of the apartment). For each scene, the following screen area categories were defined: (i) A (visible) referring entity: the agent or the text box (the agent area is further subdivided into face and body areas); (ii) The reference object: the object currently described; (iii) The apartment layout area (a designated, permanent reference object); (iv) Other screen areas. A program has been written to map eye-tracking data to $xy-$coordinates of the video sequence and count the gaze points for each category. All data accounted for in the analysis are derived from subjects' left eyes.

Results of Attention Tracking

The core of our results was distilled from analyzing eye movements of subjects. The level of statistical significance was set to 5%. For multimedia presentations, similar hypotheses can be found in [2].

Focus of Attention Hypothesis: The hypothesis is tested by restriction to scenes where the referring entity (agent, text, voice) referr to some item of the apartment. An analysis of variance (ANOVA) showed that users focus on the reference objects more in the Voice version than in either of the Agent or the Text version ($F(2,9) = 8.2$; $p = 0.009$). The result for the map area, while not statistically significant, shows a tendency toward a similar distribution of gaze points ($F(2,9) = 2.8$; $p = 0.11$). Those results suggest that gaze points are not randomly distributed across the screen area but depend on the presence or absence of a visible presentation medium. When an agent or text box is present, users' attentive focus is more evenly shared between the presentation medium and the presented material.

Locked Attention Hypothesis: This hypothesis compares the portions that subjects focus on the agent or the text box. The mean for the agent is 18% of the total number of gaze points, and the mean for the text box is 32%. The t-test (one-tailed, assuming unequal variances) showed that subjects look significantly more often at the text box ($t(6) = -2.47$; $p = 0.03$). This result can be seen as evidence that users spend considerable time for processing an object that gradually reveals new information (as for the text box).

Shift of Attention Hypothesis: We performed a (preliminary) spatio-temporal analysis of eye movement data. For example, when the agent speaks the sentence "To your left is the layout of the apartment. As you can see, the apartment includes: bedroom, living room, dining room, den, kitchen and bathroom", some subjects' focus of attention was first on the agent's face, next on the layout area, then it traversed back to the agent's face, and finally shifted to the layout area. This suggests that users expect agents to provide meaningful conversational cues and hence follow the agent's verbal and non-verbal instructions.

Agent Face–Body Hypothesis: This hypothesis has been tested by summarizing gaze points which are contained in either the agent face or the agent body region. It could be shown that subjects were looking mostly at the agent's face (mean = 83.1%), which supports the hypothesis that users interact socially with synthetic agents.

Results of Affect Tracking

In order to investigate subjects' overall affective state during the presentation, their bio-signals (skin conductance and heart rate) were analyzed. However, the study did not support the hypothesis that presentations guided by different media, such as an agent, a text box, or speech only, lead to significantly different physiological signal levels.

3. CONCLUSIONS

In this paper, novel methods to evaluate the interaction of users with different types of interfaces (synthetic agent, text, voice) have been introduced, which are based on tracking eye movements and physiological information of users. Primarily, it was demonstrated that the attentive focus hypothesized from gaze points constitutes rich information about users' actual interaction behavior with computer interfaces and that users interact in a natural and social way with synthetic interface agents.

Acknowledgments

This research was partly supported by the Research Grant (1999–2003) for the Future Program of the Japan Society for the Promotion of Science (JSPS).

4. REFERENCES

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

[2] P. Faraday and A. Sutcliffe. An empirical study of attending and comprehending multimedia presentations. In *Proceedings of ACM Multimedia 96*, pages 265–275, Boston MA, 1996.

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