Modalities for Building Relationships with Handheld Computer Agents

Timothy Bickmore
Assistant Professor
College of Computer and Information Science
Northeastern University
360 Huntington Ave, WVH 202
Boston, MA 02115 USA
bickmore@ccs.neu.edu

Daniel Mauer
College of Computer and Information Science
Northeastern University
360 Huntington Ave, WVH 202
Boston, MA 02115 USA
daniel@ccs.neu.edu

CHI 2006, April 22–27, 2006, Montreal, Canada.
ACM 1-xxxxxxxxxxxxxxxxx.

Abstract
In this paper we describe the design of a relational agent interface for handheld computers and the results of a study exploring the effectiveness of different user-agent interaction modalities. Four different agent output modalities—text only, static image plus text, animated, and animated plus nonverbal speech—are compared and their impact on the ability of the agent to establish a social bond with the user and the perceived credibility of information delivered is evaluated. Subjects generally preferred the two animated versions of the system, as well as establishing strong social bonds with them.

Keywords
Relational agent; embodied conversational agent; affective computing; social interface; handheld computers, PDAs.

ACM Classification Keywords
H5.2 [Information Interfaces and Presentation]: User Interfaces—Evaluation/methodology, Graphical user interfaces, Interaction styles, Natural language, Theory and methods, Voice I/O.

Introduction

Relational agents are computer agents designed to build and maintain long-term, social-emotional relationships with people [3]. Such relationships may be important in application domains such as education, sales and marketing, healthcare and counseling. Although these agents have been developed for desktop computers and immersive displays in which the agent is projected as a life-sized virtual person, no research has been done to date on the effectiveness and affordances of relational agents on handheld computers, such as Personal Digital Assistants (PDAs).

Handheld computers may provide an especially effective platform for relational agents. Because they are typically carried with users wherever they go, they can be accessed whenever users have a need for interaction (e.g., for health advice), potentially leading to a greater sense of reliability and trustworthiness in the agent compared to equivalent agents on desktop computers. Simply carrying the agent around may also lead to greater social bonding, due to greater contact time and closer physical proximity, giving the sense that the agent is an integral part of one's life and life experience. Since most handheld computers are not shared among users, agents running on them may also promote strong social bonding due to a greater sense of ownership and exclusivity than is possible using desktop systems, which are usually shared among household members.

We have been developing relational agents in the healthcare domain [2,3]. Specifically, we have been exploring their efficacy for health behavior change applications, such as exercise and diet promotion. In these applications, handheld computers provide another great advantage over immobile systems: when coupled with sensing devices and the user's daily calendar, they can initiate interactions with the user. For example, a smoking cessation advisor could detect when a user is lighting a cigarette and initiate a problem-solving discussion to help them stop, or an anxiety disorder counselor could initiate a deep breathing exercise just prior to a scheduled stressful event.

One significant problem in adapting these agents to handheld computers is in designing appropriate and effective interaction modalities. Most prior work in developing relational agents has focused on embodied conversational agents—animated agents that emulate face-to-face interaction using speech and nonverbal behavior [4]—as the modality of choice, given their ability to use a wide range of behaviors to display emotion and attitude. Other work on the development of automated health advisors has focused on speech-based interactions. One of our concerns in developing handheld health advisors is that users will not feel comfortable conducting speech-based interactions about their health status and behavior at work or in public environments due to privacy issues.

Thus, alternative agent interaction modalities need to be developed for handhelds that are effective at both relationship building and counseling, but that do not rely on speech. To explore this design space, we have constructed four different interfaces for a handheld relational agent —text only, static agent image plus text, animated agent, and animated agent plus nonverbal speech—and conducted a study to determine the relative effectiveness of each modality.

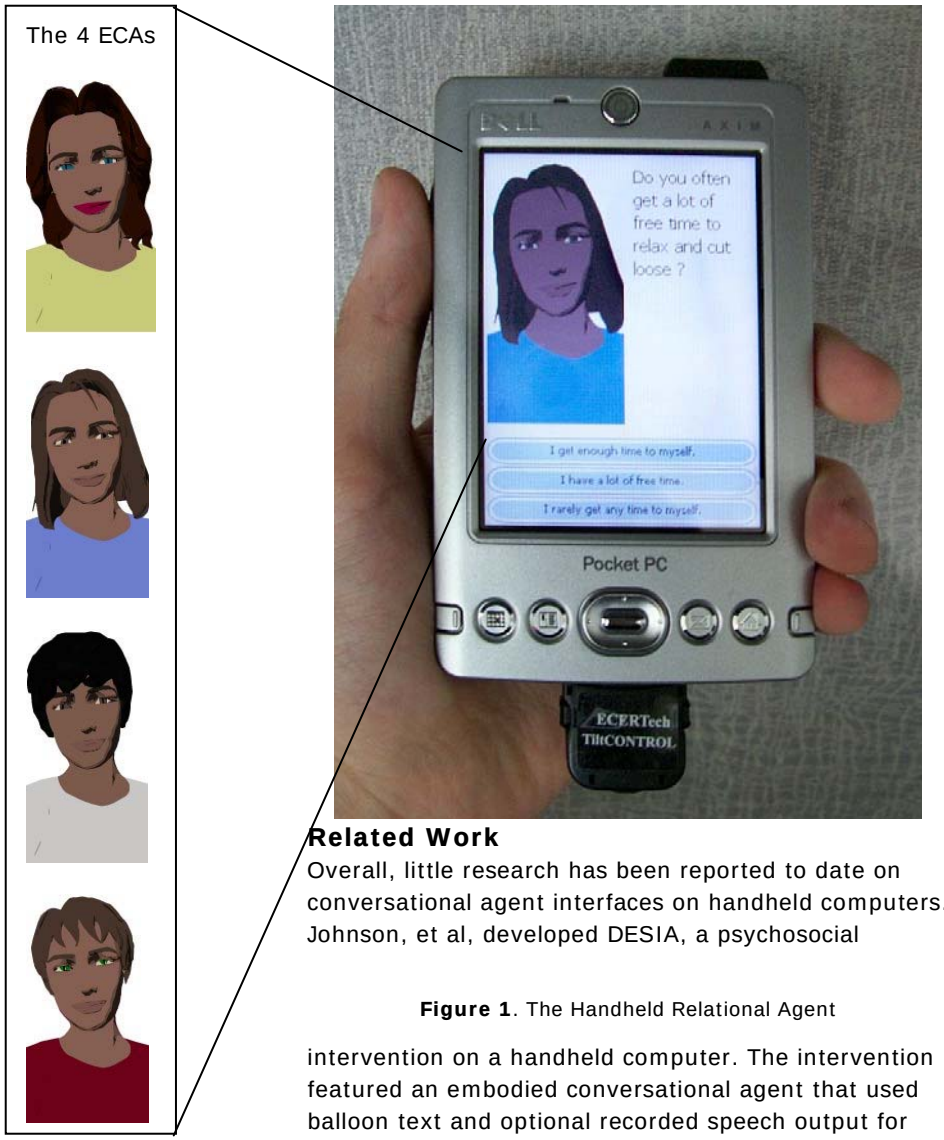

Figure 1. The Handheld Relational Agent

Related Work

Overall, little research has been reported to date on conversational agent interfaces on handheld computers. Johnson, et al, developed DESIA, a psychosocial intervention on a handheld computer. The intervention featured an embodied conversational agent that used balloon text and optional recorded speech output for the agent utterances [8]. Comparative evaluation of the different modalities (text vs. text and speech) was not reported.

A few studies have also been conducted to characterize user verbal and nonverbal behavior in their interactions with conversational agents on handhelds. Oviatt and Adams studied speech disfluencies in children talking with a handheld conversational agent and compared it to human-human conversations [9]. Bickmore conducted a study of user interactions with an embodied conversational agent on a handheld computer to characterize the nonverbal behavior people would use in these interactions [1].

A Handheld Relational Agent

We have developed a general purpose relational agent interface for use on handheld computers (see Figure 1). The animated agent appears in a fixed close-up shot, and is capable of a range of nonverbal conversational behavior, including: facial displays of emotion; head nods; eye gaze movement; eyebrow raises; posture shifts and “visemes” (mouth shapes corresponding to phonemes). These behaviors are synchronized in real time with agent output utterances. Currently, agent utterances are displayed as text with the words individually highlighted at normal speaking speed (120 words per minute) and the nonverbal behavior displayed in synchrony (this mode of synchronized display was inspired by work by Vilhjálmsson on conversational text display in avatar systems [10]). User inputs are currently constrained to multiple choice selections.

We felt that nonverbal speech, such as backchannels (“uh huh”) as well as some discourse markers (“oh”) could be used in the interaction to add to the ability of

the agent to convey emotion and attitude and to make the conversation feel more natural, but still avoid privacy issues, since it would be impossible for an overhearer to understand the conversation based solely on these sounds. Accordingly, the agent was designed to utter these sounds (recorded audio clips) at appropriate locations in the dialogue.

Interaction dialogues are scripted in a custom XML scripting language that specifies a state transition network. Each script is written by a human author, and initially consists solely of agent utterances (written in plain text), the allowed user responses to each agent utterance, and instructions for state transitions based on those responses. Once a script is written, it is pre-processed using the BEAT text-to-embodied-speech engine [5], which adds specifications for agent nonverbal behavior. In addition, each word of each utterance is processed by a *viseme generator* (based on the freeTTS text-to-speech engine) that provides the appropriate sequence of mouth shapes the agent must form in order to give the appearance of uttering that word.

After processing, while the script contains a good deal of specific instructions for the behavior of the agent, it does not fully control the agent's actions. Rather, a command embedded in the script requests a particular action to be performed in the agent's current context, which includes variables such as facial expression and posture. Certain commands, such as expression changes, can also change the current context. It is the job of the agent control system to determine what should be presented to the user based on these factors. Finally, if there are no pending action requests, an *idle action system* takes over control of the agent, randomly performing various idle behaviors (eye blinks, posture shifts, etc.).

The architecture of the run-time system on the handheld is shown in Figure 2. The actions of the system are primarily controlled by a finite state machine, which is built at run time according to the XML script. The Agent / Interface module comprises the relational agent itself (graphics, animations, audio, etc.), as well as areas for text output and user input in the form of clickable buttons. It is driven primarily by the state machine. The state machine can also accept input from sensors, such as the ECERTech's "TiltControl" accelerometer that we plan to incorporate into a future exercise promotion system.

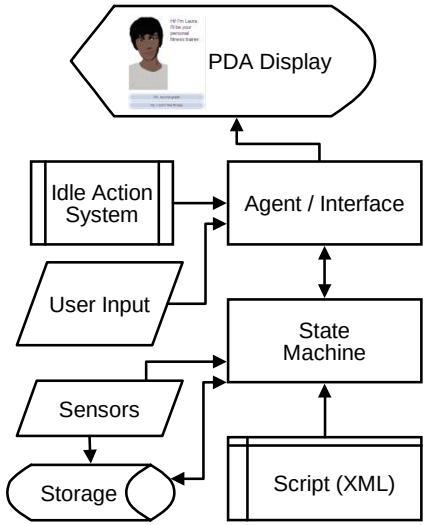

Figure 2. Software Architecture

The run-time software was developed entirely in Macromedia Flash, and we are currently using Dell Axim X30 Pocket PC computers for development and experimentation.

Comparative Evaluation Study

We conducted a study to compare four versions of the agent interface described above. For each, we assessed its relative effectiveness at establishing a social bond with the user and impacting the credibility of the information delivered, as well as user acceptance of each. The four versions evaluated were: (FULL) the full version of the animated interface (animation, text and sounds); (ANIM) the animated interface without the nonverbal speech; (IMAGE) the interface showing only a static image of the character; and (TEXT) the interface without any character.

Four structurally-similar dialogue scripts were also developed, each lasting approximately five minutes in duration. The dialogues consisted of mostly relational content (social dialogue, humor, meta-relational dialogue, etc.) but with a health tip delivered towards the end of the interaction. Four characters were also developed, based on a pre-study ranking of 14 candidate designs, and each was given a unique name.

The study has a four-condition within-subjects design, with the order of interface modes completely counterbalanced, but with a fixed order of dialogues and characters so that different modes were presented with different dialogues and characters for each subject.

Measures

Measures include: the bond subscale of the Working Alliance Inventory [7], to assess social bond; a six-item instrument to assess the credibility of the health information provided [6], and several questions about acceptability of the system and additional user attitudes towards the agent.

Procedure

Twelve subjects were recruited from the Northeastern University campus (8 males and 4 females, aged 19-21), and were compensated for their time. Subjects were given each version of the system on a separate PDA, in turn, and asked to conduct the five-minute interaction with the agent. Following each interaction the questionnaires rating the agent were administered.

Results

Data was analyzed using SPSS GLM Repeated Measures. In general, subjects preferred the two animated versions of the interface, with several measures statistically significant. There were significant differences between conditions on social bond scores (as rated on the Working Alliance Inventory, $p=.008$) and several other measures (Table 1).

Conclusion

We found that users establish stronger social bonds with handheld relational agents that are embodied and animated, compared to alternative modalities. We have several ideas for additional interaction modalities to evaluate. Synthesized or recorded speech may be usable with an earphone to avoid privacy concerns (we initially thought this would be too inconvenient for users, but the use of a wireless headset may make this workable). Another possibility is to use a low pass filter

	TEXT	IMAGE	ANIM	FULL
WAI*	3.51	3.61	4.19	4.03
PERSONAL*	2.83	3.92	4.42	4.33
SATISFACTION	3.92	3.58	4.08	4.17
CARING*	2.67	2.67	3.58	2.92
CREDIBLE	4.74	4.66	4.82	4.79
COMFORT	4.25	4.25	4.25	4.67
CONTINUE	2.75	2.83	3.42	3.17

Table 1. Primary results from study: Working Alliance Inventory (WAI) scores; rating of how PERSONAL the agent was; SATISFACTION with the system; perception of CARING by the agent; credibility of information (CREDIBLE); COMFORT with conducting this kind of interaction in a work environment; and desire to CONTINUE working with the system. Significant differences are highlighted (*).

on speech to produce muffled output that provides more affective information than the nonverbal speech we are using but still results in audio that cannot be understood by overhearers. Finally, we plan to integrate the accelerometer and conduct randomized trial on efficacy of an exercise advisor that can initiate conversations with users.

Acknowledgements

Thanks to Daniel Schulman and Ishraque Nazmi for their help on this project, and to Francisco Crespo for his assistance in conducting the study. Thanks also to Jennifer Smith for her many helpful comments on this paper. This work was supported by grant R21 LM008553 from the NIH National Library of Medicine.

References

[1] Bickmore, T., Towards the Design of Multimodal Interfaces for Handheld Conversational Characters, *CHI'02*, 2002, pp. 788-789.

[2] Bickmore, T., Caruso, L., Clough-Gorr, K., and Heeren, T. "It's just like you talk to a friend" - Relational Agents for Older Adults. *Interacting with Computers,* to appear).

[3] Bickmore, T. and Picard, R. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transactions on Computer Human Interaction,* 12, 2, (2005) 293-327.

[4] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., Eds., *Embodied Conversational Agents,* The MIT Press, Cambridge, MA, 2000.

[5] Cassell, J., Vilhjálmsson, H., and Bickmore, T., BEAT: The Behavior Expression Animation Toolkit, *SIGGRAPH '01*, 2001, pp. 477-486.

[6] Fogg, B., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., and Brown, B., Web Credibility Research: A Method for Online Experiments and Early Study Results, *ACM CHI 2001*, 2001, pp. 295-296.

[7] Horvath, A. and Greenberg, L. Development and Validation of the Working Alliance Inventory. *Journal of Counseling Psychology,* 36, 2, (1989) 223-233.

[8] Johnson, W., LaBore, C., and Chiu, Y., A Pedagogical Agent for Pyschosocial Intervention on a Handheld Computer, *AAAI Fall Symposium on Dialogue Systems for Health Communication*, 2004.

[9] Oviatt, S. and Adams, B., *Designing and Evaluating Conversational Interfaces with Animated Characters.* In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds., *Embodied Conversational Agents,* MIT Press, Cambridge, MA, 2000, pp. 319-345.

[10] Vilhjálmsson, H., **Avatar Augmented Online Conversation,* Media Arts & Sciences, MIT, Cambridge, MA, 2003.