Time to go ONLINE! A Modular Framework for Building Internet-based Socially Interactive Agents

Mihai Polceanu∗
polceanu@enib.fr
ENIB
Brest, France

Christine Lisetti
lisetti@cis.fiu.edu
Florida International University
Miami, USA

ABSTRACT

Although socially interactive agents have emerged as a new metaphor for human-computer interaction, they are, to date, absent from the Internet. We describe the design choices, implementation, and challenges in building EEVA, the first fully integrated platform-independent framework for deploying realistic 3D web-based social agents: with real-time multimodal perception of, and response to, the user’s verbal and non-verbal social cues, EEVA agents are capable of communicating rich customizable content to users in real time, while building and maintaining users’ profiles for long-term interactions. The modularity of the EEVA framework enables it to be used as a testbed for agents’ social communication model development of increasing performance and sophistication (e.g. building rapport, expressing empathy). We discuss a case study in which we show how we used the EEVA framework to create dialog content for a health agent to deliver an online tailored behavior change health intervention, and we show its feasibility by analyzing the response time of the system over the Internet.

CCS CONCEPTS

• Computing methodologies → Intelligent agents; • Computer systems organization → Real-time system architecture; • Applied computing → Health care information systems.

KEYWORDS

web-based 3D character; multimodal interaction; real-time virtual counseling

ACM Reference Format:

1 MOTIVATION

As human-computer interaction (HCI) has become increasingly present in daily life contexts involving socio-emotional content (e.g. medicine, education, entertainment), socially interactive virtual agents – also known as Embodied Conversational Agents (ECA) or as Intelligent Virtual Agents (IVA) – have emerged over the past decade as a new metaphor for HCI to address users’ need for natural interfaces simulating human-human conversations.

Building an IVA, however, is no easy feat and presents many interdisciplinary challenges. Whereas having socially appropriate interactions can be challenging even for humans at times, generating artificial social behaviors requires a mix of technology, psychology and art. Indeed, social appropriateness during dialogues, requires (apart from choosing an appropriate topic) knowing how to use

∗Current affiliation: University of Greenwich, UK.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

IVA ’19, July 2–5, 2019, PARIS, France
© 2019 Copyright held by the owner/author(s).
https://doi.org/10.1145/3308532.3329452
different channels of communication to establish and maintain rapport via verbal and non-verbal cues [6, 23], such as respectful eye contact [7], motor mimicry and synchronorous postures [3], expression of facial and other nonverbal social cues that are congruent with verbal utterances and emotional states, among others.

In spite of such complexity, IVA researchers have leveraged latest progress in affective computing [2, 4, 24] to build agents with subtle social cues and responses [1, 7, 14, 16, 16, 20, 25]. IVAs are becoming able to establish some rapport [8], express (some) empathy [9, 11, 15].

In spite of their success, however, IVA development did not scale with the new ubiquitous connected devices and latest progress on 3D graphics that can be rendered on internet browsers.

Whereas a few attempts have been made to build web-based 3D ECAs [12, 18, 22], their implementation is still very rudimentary, and none provide an integrated framework for web-based IVA development, including social cues modeling and dialog generation.

In health care, where human personnel are vastly outnumbered by people who need aid, virtual health agents (also referred to as virtual health coaches) capable of screening or providing empathic support to individuals, anytime anywhere, about their lifestyles (e.g. alcohol, drug or nicotine consumption, exercise or lack of; eating habits) have not only been found promising by healthcare research, but also better accepted by users than text-only computer-based interventions [11]. Other health-related agents have been emerging [5, 10, 19, 21], but their lack of availability on the web diminishes their potential impact.

In order to be effective, a health agent needs to be easily accessible (via common communication devices, at any time), usable (have an easy to use interface), enjoyable (provide a positive user experience), responsive to user’s emotional behaviors (establish and maintain rapport) and scalable (accommodate an increasing number of users without computational overhead).

2 CASE STUDY: SOCIALLY INTERACTIVE HEALTH AGENT

The EEVA framework (Figure 1) is a cross-platform system (Fig. 2) that can be used to develop a web-based ECA capable of delivering behavior change health interventions. The current case study consists in delivering a brief motivational interviewing (BMI) intervention for at-risk behaviors such as alcohol consumption, over eating, smoking, lack of exercise.

As detailed in [13], the content of any BMI is clearly structured into a sequence of four steps, in addition to an initial greeting and a closing statement (with potential referral of resources for healthy lifestyles) [13]: 1) screening the person’s lifestyle with a series of questionnaires; 2) providing normative feedback about the person’s lifestyle; 3) if the person is found to have lifestyle patterns placing them at risk (as determined in steps 1-2), assessing what level or readiness to change the at-risk behavior(s) the person is experiencing (from not at all, to unsure, to ready); and 4) collaborating with the person to create a behavior change plan that is aligned with the level of readiness determined in step 3.

Each step has a number of questions that prompt the user to input one answer, multiple choices, or typed or spoken natural language. The system output consists in the feedback given by the virtual character, along with visual content such as text, images, videos and HTML. The framework provides a set of authoring tools created to facilitate the creation of content for diverse use cases. The interface allows the content creator to input multimedia (visual and audio content) relevant to a particular scenario that is to be executed by the virtual character.

The EEVA framework also enables the creation and retrieval of user models that can be used to tailor and personalize the interaction with the ECA. The user’s answers to the various questionnaires are saved to provide personalized feedback – in our case, normative feedback about their [un/healthy] lifestyle. The user model consists in storing the user’s answers to the agents’ utterances, along with a set of scores that are calculated based on these results. These scores allow scenario branching, using conditional guards.

2.1 Runtime Evaluation over the Web

To evaluate our framework, we deployed it on a custom built server running Ubuntu 14.04 LTS with a Quad-core 2GHz CPU and 4GB RAM and measured the response time of critical system components on connections between two continents (see Table 1). The server was configured with SSL/TSL encryption, in order to satisfy the WebRTC security standards for accessing the user’s camera and microphone. This also permitted a smoother user experience, as during non-encrypted connections the user consent must be asked for each input device reactivation (standard security policy).

We tested two main types of network connections available to the public: broadband and 4G mobile data. All experiments were performed under “first run” conditions - i.e. no cache mechanism was used, to simulate the first time a new user would connect to the system. The majority of the launch time consists in loading the 3D visual data (the 3D character and surrounding scene), taking on average 30 and 25 seconds on 4G and Broadband respectively, while the lightweight version amounts to 16 and 10 seconds respectively. Each component is asynchronously loaded, thereby limiting the additional overhead.

While using caching techniques removes this significant load time for returning users, real time interaction is dependent on the response time of speech synthesis and recognition (among other factors). The experiment results show that our framework can maintain real-time interaction on both connection types, amounting to an average of 1 second to produce each spoken sentence and fast
continuous speech recognition using the functionality built into the Chrome browser.

Table 1: Average response time and standard deviation analysis for EEVA (in millisecond); 4G/Broadband connections over the Internet between North America and Europe; caching disabled (first run).

<table>
<thead>
<tr>
<th>Functionality</th>
<th>4G mobile data</th>
<th>Broadband internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unity 3D character</td>
<td>30018 +/- 663</td>
<td>24626 +/- 1910</td>
</tr>
<tr>
<td>Light 3D character</td>
<td>16612 +/- 1471</td>
<td>10934 +/- 2008</td>
</tr>
<tr>
<td>TTS (sentence)</td>
<td>939 +/- 381</td>
<td>551 +/- 141</td>
</tr>
<tr>
<td>TTS (word)</td>
<td>72 +/- 40</td>
<td>44 +/- 23</td>
</tr>
<tr>
<td>Speech recognition</td>
<td>~30 (Offline processing)</td>
<td></td>
</tr>
<tr>
<td>Entire HTTP request</td>
<td>1124 +/- 166</td>
<td>784 +/- 66</td>
</tr>
<tr>
<td>DOM loading</td>
<td>2313 +/- 80</td>
<td>1635 +/- 224</td>
</tr>
</tbody>
</table>

The experiments (Table 1) show that the main distributed functionalities of the EEVA framework do allow real-time interaction and acceptable loading times even for the first run. Scaling up to a large number of users does not imply a linear increase of server capacity, as after the download, most critical components run on the client-side, while only some lightweight communication with the server is performed.

3 FUTURE WORK ADDING EMPATHIC CUES

Future work will involve carrying out experiments and evaluations of nonverbal models of behavior by end-users of the health agent system, in terms of the realism of the IVA behaviors, as well as the end-users’ perceived sense of rapport with the IVA delivering the health intervention.

We will also investigate data-driven approaches to modelling the character’s behavior and deploying them in directly into the client-side environment, to be integrated with the EEVA mainframe. This brings a significant advantage: the end-users’ facial images would not need to exit the user’s personal device for the system to function, thereby removing any potential privacy concerns about sharing identifiable facial images over the network.

ACKNOWLEDGMENTS

This work has been funded by the National Science Foundation (NSF) award #1423260. Thanks to G. Ruiz, D. Rivero, S. Lunn, E. Henley, S. Bolivar for their contribution, and to anonymous reviewers for valuable feedback.

REFERENCES