Exploring the Uncanny Valley Effect with Talking Heads

Takaaki Kuratate, Kathryn Ayers, Jeesun Kim, Denis Burnham

MARCS Auditory Laboratories, University of Western Sydney, Sydney, Australia

t.kuratate@uws.edu.au, 98159215@student.uws.edu.au, j.kim@uws.edu.au, Denis.Burnham@uws.edu.au

Abstract

Here the "Uncanny Valley", where falling just short of perfection in creating synthetic humans exacts a large negative reaction, is explored with talking head animations focusing on naturalness in speech, face model and face motion. We discuss possible techniques to manipulate naturalness for each of these aspects. Outcomes of this method will provide insights for the choice of the degree of realism or naturalness required for various talking heads.

Index Terms: facial animation, uncanny valley

1. Introduction

Realistic facial animation remains a major challenge in computer graphics research. In state of the art 'photo-realistic' computer animation films, people are often critical of the animation, feeling that something is missing in the face and body motion of talking animated humans. In contrast, simpler cartoon-like computer animations typically have broad acceptance. Reaction to realistic computer-generated characters is highlighted by the "Uncanny Valley" effect, a term coined in robotics by Mori [1]. This effect involves the phenomenon that people's positive emotional response to robots increase as these are made increasingly more human-like in appearance and motion until a certain level is reached, and then suddenly these positive reactions abate (as the robots become very close to human). Mori proposed that the positive responses will be re-instated if the robot's appearance and movement improve and become almost indiscernible from humans. In this paper, we will present methods to explore this Uncanny Valley effect with talking head animations. The aim is to uncover the critical factors required to achieve a human-friendly face animation in which the focus is on auditory-visual aspects of talking heads.

2. Auditory-Visual Stimuli

In this method, we manipulate the naturalness of three factors: speech, face model and face motion, as explained below.

We used recorded speech data from human subjects and two types of altered speech to create stimuli. Sine-wave speech (SWS) [2] data was synthesized from the natural speech, while completely synthetic speech data was created using a text-to-speech (TTS) system. The natural speech was used to impose timing information for the TTS-based speech. We used software called "Repeat After Me" on Mac OS X which can impose pitch and duration from recorded speech and calls Apple's TTS engine (included in Developers environment Xcode). Using the original and the two synthesized speech data sets, we applied voice morphing techniques via STRAIGHT [3] to create gradually changing speech from synthesized to natural. All speech sets used the same visual cues.

Face naturalness and animation was manipulated using a 3D talking head animation system based on principal component analysis results [4]. This system can synthesize talking faces from a small set of parameters consisting of 6 degree-of-freedom head motion data and linear combination values of principal components (PCs) of the face model. These linear combination values are mapped from recorded face motion data collected along with corresponding natural speech data (used as described above) from human subjects. This animation system can animate any 3D human face model from a single posture using a 3D face database, can use various non-human face models such as statues, dolls and stuffed-animals, and can 3D morph between any models. These features allow us to easily select and synthesize different degrees of face model realism for our experiments.

Controlling face motion realism was accomplished as follows. Since the face in each frame was synthesized as a linear combination of PCs extracted from various static 3D facial expressions including speech-related mouth postures, changing the number of principal components resulted in statistically "degraded" face motion (e.g., using the first PC synthesizes face motion from only jaw-related deformation). We also modified the temporal aspects of face motion. The linear combination values are highly correlated with the original 3D trajectories of face makers. Thus modifying the linear combination trajectories provided different naturalness of face motion (e.g., make it smoother, or more jagged).

3. Conclusion

The above method provides a basic framework within which to evaluate the impact of speech, movement and appearance factors on the Uncanny Valley effect by using perceptual experiments. Such perceptual experiments will provide a guide to the respective grain size and boundary conditions over which the perception of naturalness plays out.

4. Acknowledgements

We acknowledge Australian Research Council Discovery (DP0666891), and Special Initiatives (TS0669874) support.

5. References

[1] Mori, M. "The Uncanny Valley", http://www.androidscience.com/theuncannyvalley/proceedings2005/uncannyvalley.html, accessed on 15 Apr 2008.

[2] Remez, R. E., & Rubin, P. E., "On the perception of intonation from sinusoidal sentences". Perception & Psychophysics, 35:429-440, 1984.

[3] Kawahara, H,. & Matsui, H., "Auditory morphing based on a elastic perception distance metric in an interference-free time-frequency representation", Proc. ICASSP'03, 256-259, 2003.

[4] Kuratate, T., "Statistical analysis and synthesis of 3D faces for auditory-visual speech animation", Proc. AVSP'05, 131-136, 2005.

Copyright © 2008 ISCA