Immersion in Multimodal Gaming: Playing *World of Warcraft* with Voice Controls

Antonio Ricciardi
Stanford University
353 Serra Mall
Stanford, CA 94305, USA
aricciardi@stanford.edu

Jae min John
Stanford University
353 Serra Mall
Stanford, CA 94305, USA
jaemin@cs.stanford.edu

ABSTRACT

Recent trends in video game hardware design suggest that the traditional game controller is growing out of fashion. As designers begin to explore new, multimodal control schemes, with both touch and non-touch inputs, it is important that they understand how this affects the player's experience. To discover how such interfaces influence the immersiveness of a game, we compared player responses using two multimodal and two unimodal control schemes for Blizzard Entertainment's *World of Warcraft*. Voice and keyboard inputs were used for two different tasks, and the four possible combinations were compared using a within-subjects design. Immersion was measured quantitatively with eye-tracking and qualitatively using surveys. The results suggest that the amount of immersion a player experiences in a game is related to the amount of cognitive effort required by the control scheme. In our study, players were most attentive while using a multimodal interface which required a medium amount of effort. The implications of our results are that video game designers must optimize the cognitive load that a game imposes on a player by choosing the right combination of modalities and correctly mapping those modalities to controls.

ACM Classification: H.5.2 [Information Interfaces and Presentation] User Interfaces - *Input devices and strategies.*
General terms: Design**,** Human Factors

Keywords: Multimodal, immersion, games, controls.

INTRODUCTION

For the past few decades, players of computer and home console video games have been limited to inputs with a mouse and keyboard or a button-based controller. However, these games have recently begun to incorporate inputs using modalities other than touch, such as voice recognition. An important question in the future of the video game industry is how these new control schemes will affect the player's experience.

For most video games, one of the designer's main goals is to immerse the player in a virtual world. It would therefore be advantageous to know whether certain game interfaces and control schemes promote a greater sense of immersion than others. Previous work has shown that multimodal game outputs, such as auditory and visual cues, can have a positive effect on immersion [2]. However, little research has been done involving multimodal player input. The goal of our study was to determine how a multimodal control scheme affects the immersiveness of a video game.

EXPERIMENT

To answer our question, we modified the user interface of Blizzard Entertainment's *World of Warcraft* to support two multimodal and two unimodal control schemes. We then compared these control schemes using qualitative and quantitative measures of immersion.

Each of two different game tasks (controlling your character and communicating with your teammates) was assigned two possible inputs: voice and keyboard. Using the *World of Warcraft* API, we implemented a voice recognizing system which uses voice commands to trigger in-game character actions. For instance, when a player would say "lightning bolt", the character onscreen would cast a bolt of lightning at an enemy. Using some other default features of the game, we produced four different conditions: voice control with voice chat (VV), voice control with keyboard chat (VK), keyboard control with voice chat (KV), keyboard control with keyboard chat (KK).

Hypothesis

The use of multimodal inputs in an interface has been shown do decrease a user's cognitive load by increasing the amount of working memory available [3]. Furthermore, it is believed that initial engagement in a game requires a certain cognitive investment from the player, creating a barrier for immersion [1]. Consequently, we hypothesized that a player would become more immersed in a game with a multimodal control scheme (KV or VK) than with a unimodal one (VV or KK).

Procedure

We ran 10 subjects having little or no experience with *World of Warcraft*, using a within-subjects design. The subjects played using each control scheme for 10 minutes in a randomized order, separated from their teammate (one of the experimenters) by an opaque screen. After each condition, the subjects took a short survey which measured their levels of challenge, focus, involvement and enjoyment. The survey served the additional purpose of preventing immersion from carrying over between conditions. We also recorded the subjects' faces to track their eye movements.

Results

From the eye-tracking data, we calculated a 3-D gaze vector using the subjects' vertical and horizontal gaze angles. We then calculated and normalized each subject's mean "fixation duration" based on the number of frames during which his or her gaze vector did not change by more than 20 degrees. Fixation duration has been previously shown to correlate with immersion in video games [4].

This data showed a significant increase ($p < 0.05$) in mean fixation duration with condition VK compared to other conditions. The other multimodal interface (KV) showed no significant difference from the unimodal conditions.

Unfortunately, the subjective data was inconsistent with the objective data. Although it had a low significance ($p > 0.2$), the survey data showed higher values in all four categories – challenge, focus, involvement and enjoyment – for the VV group, followed closely by KV.

CONCLUSIONS

We believe the inconsistencies between the subjective and objective measures of immersion were due in large part to the subjects' conscious preference for voice chat, which may have been induced by flaws in our experimental design. Specifically, we were collocated with the subjects and knew many of them on a personal basis. As a result, the survey responses strongly favored VV and KV.

From the above eye-tracking results, it is clear that our changes to the game's user interface had an impact on the subjects' play experience. However, our hypothesis that using a multimodal control scheme would increase a player's focus was only half-correct; only one of the multimodal interfaces (VK) had a significantly higher mean fixation duration than the unimodal versions.

While these results may have a number of possible explanations, we believe they are strongly related to variations in the amount of cognitive load imposed on a player by the different control schemes. On the one hand, using a unimodal interface limits the amount of working memory available to the player, increasing the effort required to achieve immersion. On there other hand, using the KV interface (with voice chat) lowered this cognitive barrier *too much*. As a result, the players experienced underload and grew distracted, since the game did not require a large amount of their attention (Figure 1). Cognitive underload has been shown to significantly reduce user attention, especially when using highly automated interfaces [5]. We believe the difference between the two multimodal groups was caused by the fact that voice chat is significantly easier than keyboard chat in *World of Warcraft*.

An important consequence of these results is that designers cannot simply increase the immersiveness of a game by increasing the number of modalities used by a control scheme; they must find the correct assignment of modalities to tasks in order to optimize the player's cognitive load.

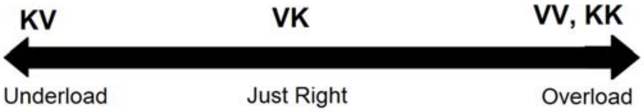

Figure 1: The relationship between cognitive load and control scheme in *World of Warcraft*.

FUTURE WORK

Before any work is done to build off this study, it is important that these results are reproduced using a more formal experimental design. Specifically, a between-subjects design with more subjects would prevent any bias from occurring between conditions. Teammates should also be unfamiliar with each other and located in separate rooms.

Once this is done, the next step is to investigate different methods for optimizing cognitive load. This means learning how to determine the right combination of modalities and how to assign those modalities to controls. One problem which was evident from our results is that, although keyboard chat was more successful in terms of immersion, subjects consciously preferred voice chat because it was more natural. Thus, the optimization of cognitive load must be done within a certain set of constraints so as not to sacrifice other aspects of the player's experience.

ACKNOWLEDGMENTS

Thanks to Maria Jabon of Stanford's Virtual Human Interaction Lab, as well as Prof. Scott Klemmer and Joel Brandt for putting together such a great class.

REFERENCES

1. Brown, E. and P. Cairns. A Grounded Investigation of Game Immersion. In *CHI '04 extended abstracts on Human factors in computing systems*, CHI, Vienna, 2004.

2. Nesbitt, K. V. and I. Hoskens. Multi-sensory Game Interface Improves Player Satisfaction but not Performance. *Conferences in Research and Practice in Information Technology*, Vol. 76, AUIC, Wollongong, 2008.

3. Oviatt, S. Multimodal Interfaces. In *The Human-Computer Interaction Handbook*, A. Sears, J. Jacko, ed., Lawrence Erlbaum, 2003, pp. 286-304.

4. Tijs, T. J. W. Quantifying Immersion in Games by Analyzing Eye Movements. Department of Computer and Systems Science, Royal Institute of Technology, Stockholm, 2006.

5. Young, M. S., and N. A. Stanton. Attention and Automation: New Perspectives on Mental Underload and Performance. *Theor. Issues in Ergon. Sci.*, 2002, Vol. 3, No. 2, pp. 178-194.