Real Time Gesturing in Embodied Agents for Dynamic Content Creation

Demonstration Track

Hazel Watson-Smith
Soul Machines
Auckland, New Zealand
hazel.watson-smith@soulmachines.com

Felix Marcon Swadel
Soul Machines
Auckland, New Zealand
felix.swadel@soulmachines.com

Jo Hutton
Soul Machines
Auckland, New Zealand
jo.hutton@soulmachines.com

Kirstin Marcon
Soul Machines
Auckland, New Zealand
kirstin.marcon@soulmachines.com

Mark Sagar
Soul Machines
Auckland, New Zealand
mark.sagar@soulmachines.com

Shane Blackett
Soul Machines
Auckland, New Zealand
shane.blackett@soulmachines.com

Tiago Ribeiro
Soul Machines
Lisbon, Portugal
tiago.ribeiro@soulmachines.com

Travers Biddle
Soul Machines
Auckland, New Zealand
travers.biddle@soulmachines.com

Tim Wu
Soul Machines
Auckland, New Zealand
tim.wu@soulmachines.com

ABSTRACT

The content creation industry is experiencing significant growth and the utilisation of Real-Time Gesturing in embodied agents presents an excellent opportunity to enhance the communication of text. Using the proposed system, raw text can be parsed in real-time and an appropriate emotional and gestural performance is generated. It can also be configured to convey personality traits using elements such as emotional state and responses to stimuli, gesture rate, type, size and speed, and augmented with inserted markup tags.

KEYWORDS

Autonomous Animation; Content; Creator; Gestures; Real-Time

1 INTRODUCTION

Embodied Agents with Real-Time Gesturing present an excellent opportunity to enhance the communication of written content and unlock a new paradigm of content creation. In 2021, the global market for digital content creation reached a value of an estimated $12.2 billion USD. The industry is expected to reach a value of around $24.73 billion USD by 2027 [5]. With Soul Machines’ Real-Time Gesturing System, written content can be easily transformed into high-quality interactions.

2 SYSTEM DETAILS

2.1 Real-Time Gesturing Computation

Plain text is taken as input and is broken down into a clause tree, which defines the lemma of each word, and which part of speech it fulfills. Each word is evaluated for importance, emotional, and symbolic content, to inform the placement of vocal emphasis, beat gestures, facial expressions and symbolic gestures. Finally, the overall sentiment of the sentence is evaluated. Modules within the Real-Time Gesturing system each propose gestures and expressions to perform. Where conflicts arise, a resolver module chooses which gestures to retain based on priority configuration. This enables the Embodied Agent to perform appropriate gestures and facial expressions to accompany and enhance the text and its communication. The facial and body animation systems work in conjunction with...
a real-time lip sync system [9]. For example, the text “The sooner you open up and become all-inclusive, the sooner we have everything.” The words ‘open’, ‘all-inclusive’ and ‘everything’ would be candidates for wide, expansive gestures and may be included in the final gesture string.

2.2 Stylization and Affect-driven performance

The gestural and emotional performance is easily configurable to create a desired style of behaviour with specific target traits, such as extroversion, openness, and agreeableness, or more physical traits like handedness. This is achieved by adjusting parameters such as emotional responses, as well as gesture selection, size and rate. Once this is configured, it can be saved as a preset style of behaviour that can be reused to consistently create characteristic performances.

The emotional performance is connected to a neurobehavioral model [6–8], creating a connection between the affective content of the text with autonomous facial and body animation. The internal emotional state of the Embodied Agent influences their gestural performance to better suit the spoken content.

All of this is computed in real-time and with no code implementation on the user side. There are opportunities to pair this technology with generative language models to bring these conversations to life.

Figure 1: Real-Time Gesturing System pathway from text input to Embodied Agent Model

While the Real-Time Gesturing system allows you to parse in raw text and receive a full dynamic emotional and gestural performance, a non-technical user, such as a scriptwriter or director, can augment the performance with simple tags in the text, such as #Smile. This instructs the Embodied Agent to perform the specified gesture or expression in that specific place. This allows for further control and the addition of facial expressions or gestures that are not automatically triggered by the text, such as conveying sub-text, sarcasm and brand-specific messaging.

In the context of content creation, Real-Time Gesturing generates a solid foundation of animation that can be refined and enhanced by the use of tags in small sections of the text. This allows the creators to focus on key moments of the performance rather than laboriously marking up every detail of every moment. This gives a more practical level of control to the creator, akin to the control a director would have with an actor and a more natural end performance.

3 DEMO

The video demo features a conversation with an embodied agent demonstrating Real-Time Gesturing. Apart from the waves and thumbs up, which were inserted as markup tags, the conversation script is plain text.

https://youtu.be/j1FR0rSkRL4

The live demo features a generative conversation connected to OpenAI’s GPT-3, with personality-guided prompts and a custom knowledge base.

ACKNOWLEDGMENTS

The authors would like to thank Amy Wang, Emma Perry, Hannah Clark-Younger, Jiali Xu, Katrina Bennett, Pavel Sumetc, Utkarsh Saxena, Xueyuan Zhang and all who contributed to this research.

REFERENCES