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

ACM Reference Format:
Hazel Watson-Smith, Felix Marcon Swadel, Jo Hutton, Kirstin Marcon, Mark Sagar, Shane Blackett, Tiago Ribeiro, Travers Biddle, and Tim Wu. 2023. Real Time Gesturing in Embodied Agents for Dynamic Content Creation: Demonstration Track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 2 pages.

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.

A dynamic, natural and semantically informed gestural and emotional performance, containing various types of arm and hand gestures including symbolic, iconic and beat gestures, and also facial, head-motion and postural performance, is generated in real-time to bring your words to life.

Rule-based gesture generators, such as BEAT [1] apply rules to generate gestures, paired with features of the text. This results in repetitive and robotic gesturing, which is difficult to customize on a granular level. Large databases of rules and gestures are required, and expensive to acquire or build. Speech-driven gesture generators, such as MoGlow [3] use neural networks to automatically generate movements from learnt gesture and speech combinations. However, these generators often work in a black-box manner and assume a general relationship between the input speech and output motion, which is not always the case. Other notable work in this field include the proceedings of the GENEA Workshop 2022 [10], Smartbody + Cerebella and the Virtual Human Toolkit [2], and Gesticulator [4].

In this demo, we present a Real-Time Gesturing system for Embodied Agents in the context of dynamic content creation.

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 fulfils. 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

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

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.

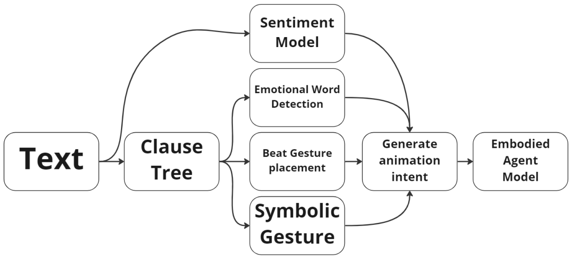

Figure 1: Real-Time Gesturing System pathway from text input to Embodied Agent Model

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/j1F0R0SkRL4

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

[1] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. *BEAT: the Behavior Expression Animation Toolkit.* Springer Berlin Heidelberg, Berlin, Heidelberg. 163–185 pages. https://doi.org/10.1007/978-3-662-08373-4_8

[2] Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All Together Now. In *Intelligent Virtual Agents*, Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 368–381.

[3] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. MoGlow: Probabilistic and Controllable Motion Synthesis using Normalising Flows. *ACM Transactions on Graphics* 39, 4 (2020), 236:1–236:14. https://doi.org/10.1145/3414685.3417836

[4] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Simon Henter, Gustav Eje abd Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A Framework for Semantically-Aware Speech-Driven Gesture Generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. ACM. https://doi.org/10.1145/3382507.3418815

[5] Expert Market Research. 2022. Global Digital Content Creation Market: By Component: Tools, Services; By Content Format: Textual, Graphical, Video, Audio, Others; By Deployment Type; By Enterprise Size; By End Use Industry; Regional Analysis; Historical Market and Forecast (2018-2028); Market Dynamics; Competitive Landscape; Industry Events and Developments. Retrieved 3/2/2022 from https://www.expertmarketresearch.com/reports/digital-content-creation-market

[6] Mark Sagar, David Bullivant, Oleg Robertson, Pauland Efimov, Khurram Jawed, Ratheesh Kalarot, and Tim Wu. 2014. A Neurobehavioural Framework for Autonomous Animation of Virtual Human Faces. In *SA '14: SIGGRAPH Asia 2014 Autonomous Virtual Humans and Social Robot for Telepresence.* ACM. https://doi.org/10.1145/2668956.2668960

[7] Mark Sagar, Paul Robertson, David Bullivant, Oleg Efimov, Khurram Jawed, Ratheesh Kalarot, and Tim Wu. 2015. BL: A Visual Computing Framework for Interactive Neural System Models of Embodied Cognition and Face to Face Social Learning. In *Unconventional Computation and Natural Computation.* Springer. https://doi.org/10.1007/978-3-319-21819-9_5

[8] Mark Sagar, Mike Seymour, and Annette Henderson. 2016. Creating Connection with Autonomous Facial Animation. *Commun. ACM* 59, 12 (2016), 82–91.

[9] Mark Sagar, Tim Wu, Xiani Tan, and Xueyuan Zhang. 2020. Real-Time Generation of Speech Animation. https://patents.google.com/patent/WO2020152657A1 Patent No. NZ75023319, Filed Jan 27th., 2020 by Soul Machines Limited, Published Jul 30th., 2020.

[10] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022. In *The GENEA Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 736–747. https://doi.org/10.1145/3536221.3558058