Conversational Gestures: Transforming Text into Full-Body Virtual Interactions

Saif Punjwani
Georgia Institute of Technology
spunjwani3@gatech.edu

Bowen Zuo
Georgia Institute of Technology
bzuo6@gatech.edu

Micah Grimes
Georgia Institute of Technology
grimesm@gatech.edu

Larry Heck
Georgia Institute of Technology
larryheck@gatech.edu

Abstract

With the increasing prevalence of virtual environments and their significance in human-agent interactions, there's an emergent need for pseudo real-time virtual gesture systems capable of authentically replicating human gestures. This paper delves into the development of a state-of-the-art system that converts text prompts into realistic virtual gestures. Our objective is to enhance the representation of human agents in virtual spaces. We build upon existing model implementations by integrating a novel Transformer (6) based approach to optimize the input of our own proprietary allocentric dataset comprising diverse and intricate human gestures. Our approach involves a better and more efficient pipeline to translate textual input into physically plausible and contextually accurate gestures. The resulting system not only replicates gestures but also captures the nuances of human motion, contributing to more lifelike and engaging interactions in virtual environments.

1 Introduction

Virtual environments have paved the way for humans to converse with digital agents through the online medium in a manner that is more fluid and dynamic (3). With a growing emphasis on this context, we see that human-agent interactions through gestures are becoming a pivotal area of improvement. Gestures are incredibly challenging to replicate, but essential in conversations between humans and human-agents. They carry meaning and communicate emotions that are hard to grasp through speech alone. To address this gap, we propose an improved system designed to transform text prompts into realistic virtual gestures, thereby enriching the representation of human agents within these digital realms. Building upon the foundational work of Evonne Ng et al (1), which explores the potential of language models in understanding and interpreting human communication, our study ventures further into the realm of human-agent interaction. We propose an advanced approach that leverages the capabilities of Transformer-based models to processing the sequential data of our proprietary dataset, which is primarily based on a comprehensive video-graphic collection of allocentric POVs of human conversation. This data embodies a wide array of expressions and actions that are central to human communication (14), as well as extensions making this model encapsulate gestures for full-body representation (16). The model uses text as an input and focuses on capturing the attention in terms of emotion and time to map the proper text-to-gesture output (17). The integration of our improvements to the Transformer-based model in conjunction with our curated dataset, enables our system to capture the subtleties of human motion, thereby allowing enhanced communication between humans and digital agents (24). Through these developments, we contribute to the progression of virtual interactions, making them more lifelike and immersive, to work towards a future where digital agents can communicate with humans in a manner that is both authentic and deeply engaging.

2 Model Adjustments and Improvements

To enhance the generation of authentic human gestures from textual prompts, we introduce a hybrid architecture that builds upon the encoder-decoder framework (6) and incorporates elements of VQ-VAE (7), optimizing for time-sequenced gesture fidelity and overall quality.

2.1 Enhanced LM-Listener Model

Our model advances the lm-listener (1) framework, integrating the contextual strengths of Transformers with the sequential handling prowess of RNNs (8). This hybrid model is tailored to process full-body gestures, allowing a deeper understanding and translation of text prompts into dynamic gesture sequences that are aligned with the speaker's intended expression and context.

$$G_t = f_{LM}(T_t, H_{t-1}; \theta) \quad (1)$$

In this formula, G_t signifies the gesture output at time t, T_t is the textual input, H_{t-1} is the preceding hidden state, and θ represents the model's parameters.

2.2 Advanced Sequence Mapping

We have developed an advanced sequence-to-sequence mapping technique that employs a dual-attention mechanism. This method ensures that the model's output gestures are not only temporally coherent but also contextually synchronized with the given text (15).

$$A_t = \alpha_{text}(T_t) + \alpha_{hist}(G_{t-1})$$

Here, A_t represents the attention-weighted output, with α_{text} and α_{hist} being the attention functions for the textual content and gesture history, respectively. This approach improves upon the traditional lm-listener model by encompassing full-body gesture dynamics.

2.3 Gesture Generation with Generative Models

A pivotal component of our model is the generative system G, which crafts the final gesture output (5) from the integration of the current text input, historical context, and previously generated gestures. Our full gesture output then becomes:

$$G_t = G(W_{history}, T_t, M_{1:t-1}) + f_{LM}(T_t, H_{t-1}; \theta)$$

Where G_t is the gesture at time t, $W_{history}$ is the sequence of words spoken before time t, T_t is the current text token, $M_{1:t-1}$ is the sequence of past generated motions with [1].

The generative system is adept at producing complex gesture sequences that are expressive and varied, trained to capture the nuances and emotional context of spoken language (11). By learning from a rich dataset, the generator can ensure that the produced gestures are not just are both accurate and natural similar to human movement.

3 Dataset and Evaluation

3.1 Dataset Composition and Improvements

Our research is anchored in the development of a robust allocentric-specific modeling framework, advancing the work on listener motion generation. We have substantially enriched our dataset to overcome the constraints of previous studies and to embed a wider spectrum of communicative contexts. The dataset, curated meticulously from various sources such as talk shows, podcasts, TED talks, and other public speaking forums, offers a multifaceted array of human gestures and expressions (21; 19). This extensive collection is crucial for training our model to recognize and replicate the nuanced spectrum of human nonverbal communication.

3.2 Dataset Preparation and Annotation

The preparation of our dataset involved a meticulous process of mapping gestures to their corresponding textual prompts. Utilizing OpenPose, an advanced keypoint detection tool, we analyzed each video frame to construct an $N \times M$ matrix. This matrix encapsulates the spatial configuration of keypoints — essentially the coordinates of significant body joints and facial landmarks.

To synchronize the gesture data with the spoken text, we combined the keypoints matrix with the transcribed speech. The resulting composite data were then flattened into a sequence format that our model could effectively process. This procedure ensures that each gesture is contextually linked to the corresponding textual prompt, allowing for a more coherent and meaningful interpretation of gestures in relation to spoken language.

4 Results

Our comprehensive evaluation demonstrates the effectiveness of our approach in generating full-body conversational gestures from text (15). Through analysis, we have established that our method has potential to outperform baseline metrics, offering a promising avenue for enhancing human-agent interactions in virtual environments.

4.1 Discussion

Initial experiments with our model demonstrate its ability to accurately generate full-body gestures from text that are both realistic and contextually synchronized with spoken narratives. This advancement suggests a promising direction for enhancing avatar realism (23) in virtual settings, reducing the dependence on complex multimodal inputs.

The model's performance, particularly in reflecting authentic listener responses, indicates its applicability in creating virtual agents with nuanced emotional resonance.

5 Limitations

While our model demonstrates promising results, it is essential to acknowledge its limitations. The current implementation focuses primarily on allocentric data sources, which while diverse, does not encompass the full range of human gestures and expressions across different cultures and contexts. Future work could expand the dataset to include more varied sources, potentially enhancing the model's versatility and applicability across a broader spectrum of virtual interaction scenarios.

6 Ethics Statement

The ethical considerations of our research are twofold. Firstly, while our model aims to enhance virtual interactions, it is imperative to consider the privacy and consent issues related to using real-life video data for training purposes. Secondly, the potential for misuse of realistic virtual agents, such as in creating deepfake content, necessitates the development of robust frameworks to ensure ethical use and application of this technology (10).

7 Acknowledgements

We thank the contributors of the original datasets and models that facilitated this research, particularly the teams behind OpenPose, PyAnnote, DECA, EMOCA, and Whisper, whose tools were instrumental in our dataset preparation and model development processes.

References

[1] Evonne Ng, et al. *Can Language Models Learn to Listen? A Quantitative Analysis on Language Model Implementations for Conversational Agents*. Proceedings of the International Conference on Learning Representations, 2017.

[2] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. *Neural Discrete Representation Learning*. Advances in Neural Information Processing Systems, 2017.

[3] Zhe Cao, et al. *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 2019.

[4] Hervé Bredin, et al. *PyAnnote: A Machine Learning Toolkit for Audiovisual Indexing*. Proceedings of the International Conference on Multimedia Retrieval, 2020.

[5] Egger, Bernhard, et al. *EMOCA: Emotionally Expressive 3D Face Reconstructions*. ACM Transactions on Graphics, vol. 40, no. 4, Article 142, 2021.

[6] Vaswani, Ashish, et al. *Attention Is All You Need*. Advances in Neural Information Processing Systems, 2017.

[7] Razavi, Ali, Aaron van den Oord, and Oriol Vinyals. *Generating Diverse High-Fidelity Images with VQ-VAE-2*. Advances in Neural Information Processing Systems, 2019.

[8] Mikolov, Tomas, et al. *Recurrent Neural Network Based Language Model*. Proceedings of the 11th International Conference on Speech and Computer, 2010.

[9] Rautaray, Siddharth S., and Anupam Agrawal. *Real Time Hand Gesture Recognition System for Dynamic Applications*. International Journal of UbiComp (IJU), vol. 3, no. 1, pp. 21-31, 2012.

[10] Slater, Mel, and Sylvia Wilbur. *A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments*. Presence: Teleoperators and Virtual Environments, vol. 6, no. 6, pp. 603-616, 1997.

[11] Zhou, C., Bian, T., Chen, K. *GestureMaster: Graph-based Speech-driven Gesture Generation*. In INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India. ACM, New York, NY, USA.

[12] Chang, C.-J., Zhang, S., Kapadia, M. *The IVI Lab entry to the GENEA Challenge 2022 – A Tacotron2 Based Method for Co-Speech Gesture Generation With Locality-Constraint Attention Mechanism*. In INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India. ACM, New York, NY, USA.

[13] Subramoney, A., Khan Nazeer, K., Schöne, M., Mayr, C., Kappel, D. *Efficient Recurrent Architectures Through Activity Sparsity and Sparse Back-Propagation Through Time*. ICLR 2023.

[14] Richard, A., Zollhofer, M., Wen, Y., de la Torre, F., Sheikh, Y. *MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1173-1182.

[15] Korzun, V., Beloborodova, A., Ilin, A. *ReCell: Replicating Recurrent Cell for Auto-regressive Pose Generation*. In INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India. ACM, New York, NY, USA.

[16] Neff, M., Kipp, M., Albrecht, I., Seidel, H-P. *Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style*. ACM Transactions on Graphics, 27(1), 1-24.

[17] Bhattacharya, U., Childs, E., Rewkowski, N., Manocha, D. *Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning*. University of Maryland, College Park, MD, USA.

[18] Rebol, M., Gütl, C., Pietroszek, K. *Real-time Gesture Animation Generation from Speech for Virtual Human Interaction*. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA.

[19] Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., Lee, G. *Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity*. ACM Transactions on Graphics, 39(6), Article 222.

[20] Kaneko, N., Mitsubayashi, Y., Mu, G. *TransGesture: Autoregressive Gesture Generation with RNN-Transducer*. In International Conference on Multimodal Interaction (ICMI '22), November 7–11, 2022, Bengaluru, India. ACM, New York, NY, USA.

[21] Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., Neff, M. *A Comprehensive Review of Data-Driven Co-Speech Gesture Generation*. arXiv preprint arXiv:2301.05339.

[22] Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H-P., Pons-Moll, G., Elgharib, M., Theobalt, C. *Learning Speech-driven 3D Conversational Gestures from Video*. arXiv preprint arXiv:2102.06837.

[23] Tang, S., Chen, C., Xie, Q., Chen, M., Wang, Y., Ci, Y., Bai, L., Zhu, F., Yang, H., Yi, L., Zhao, R., Ouyang, W. *HumanBench: Towards General Human-centric Perception with Projector Assisted Pretraining*. arXiv:2303.05675v1 [cs.CV].

[24] Ruffieux, S., Lalanne, D., Mugellini, E., Abou Khaled, O. *A Survey of Datasets for Human Gesture Recognition*. In M. Kurosu (Ed.), Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 337–348. Springer International Publishing Switzerland.

[25] Saleh, K. *Hybrid Seq2Seq Architecture for 3D Co-Speech Gesture Generation*. In INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India. ACM, New York, NY, USA.

[26] Ghorbani, S., Ferstl, Y., Carbonneau, M.-A. *Exemplar-based Stylized Gesture Generation from Speech: An Entry to the GENEA Challenge 2022*. In INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India. ACM, New York, NY, USA.