Improving Co-speech gesture rule-map generation via wild pose matching with gesture units.

Ghazanfar Ali
University of Science and Technology
Seoul, Republic of Korea
aliust@ust.ac.kr

Jae-In Hwang
Korea Institute of Science and Technology
Seoul, Republic of Korea
hji@kist.re.kr

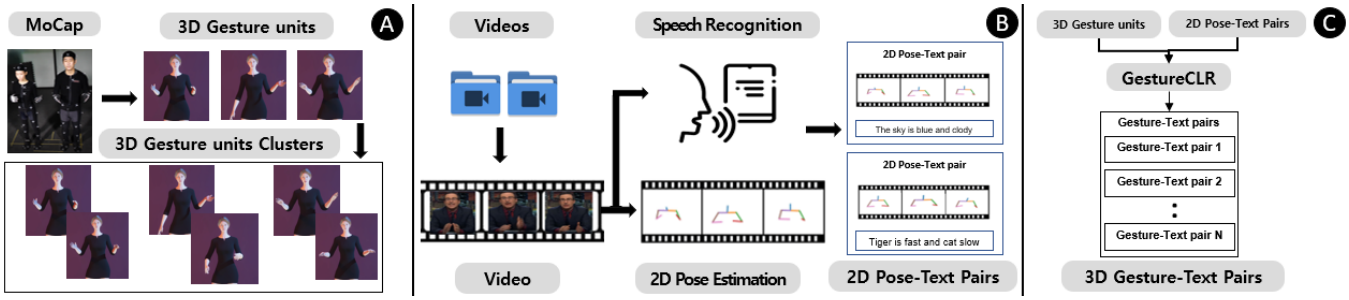

Figure 1: System overview. A) Gesture unit extraction and gesture clustering. B) 2D Pose-Text pairs extraction C) Rulemap generation by matching pose and gesture units with GestureCLR.

ABSTRACT

In this poster, we present a method to generate co-speech text-to-gesture mapping for 3D digital humans. We obtained text and 2D pose data from public monologue videos. Gesture units were obtained from motion capture sequences. The method works by matching 2D poses to 3D gesture units. We trained a model via contrastive learning to improve the matching of noisy pose sequences with gesture units. To ensure diverse gesture sequences at runtime, gesture units were clustered using K-Mean clustering. We incorporated 2035 gestures and 210k rules. Our method is highly adaptable and easy to control and use. Demo Video : https://youtu.be/QBtGdGE1Wgk

CCS CONCEPTS

• **Human-centered computing** → **Interaction paradigms**.

KEYWORDS

Virtual humans, HCI, co-speech Gestures

ACM Reference Format:
Ghazanfar Ali and Jae-In Hwang. 2022. Improving Co-speech gesture rule-map generation via wild pose matching with gesture units. . In *SIGGRAPH Asia 2022 Posters (SA '22 Posters), December 06-09, 2022.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3550082.3564185

SA '22 Posters, December 06-09, 2022, Daegu, Republic of Korea

ACM ISBN 978-1-4503-9462-8/22/12.
https://doi.org/10.1145/3550082.3564185

1 INTRODUCTION

The use of digital humans as a proxy for humans in the virtual environment is gaining popularity. Digital humans can be used in numerous ways like as guides in virtual museums[Ali et al. 2019] or actors in previsualizations[Kim et al. 2021]. Humans, along with speech, gesture to articulate their point of view. Likewise, digital humans also need co-speech gestures for natural appearances[Ali et al. 2020]. Classically it's done by motion capture and animators manually mapping the gesture with speech. However, manually mapping animations is a very laborious task. The most successful way of doing this is rule-based animations, where experts create the mapping between keywords and predefined gesture units. This method provides control and flexibility. However, this method suffers from the subjectivity of experts. Ali et al. [Ali et al. 2020] provided a way to mitigate this problem. They used the transcript and 2D pose data extracted from monologue-style public videos of human presenters. Then used, a set of predefined gesture units as sliding windows over the 2D pose sequences and extracted corresponding text for best matches. The result was an extensive text-to-gesture mapping. This method used mean cosine similarity for the pose to gesture matching, which performs low on temporal misalignment and noise. The gesture units were 54, and the rules were 84k. We used this system as a baseline system.

We present our approach that improves gesture matching via deep contrastive learning and increases the number of gestures to 2035. We added up to 210k new rules. Additionally, we used gesture clustering to diversify the output further to avoid excessive gesture repetition. This way, the output is slightly changed every time for the same input but stays relevant.

2 APPROACH

Our approach breaks the problem into three parts, i.e., 1) Gesture units and clustering, 2) 2D Pose-Text pairs extraction 3) Rulemap

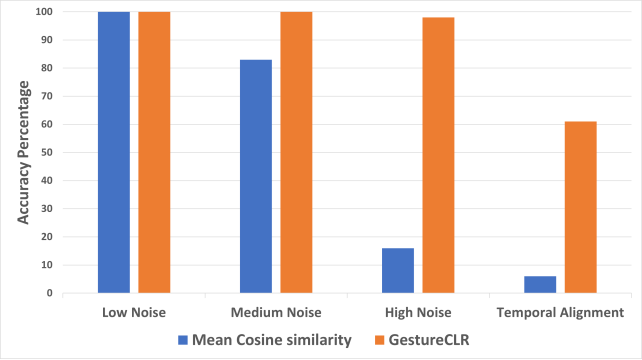

Figure 2: Comparative accuracy performance of GestureCLR and Mean Cosine Similarity.

generation with GestureCLR. We choose videos on a wide variety of topics with different presenters. The topic ranges from current affairs talk show to lectures in universities.

2.1 Gesture units and clustering

We defined a gesture unit as a short sequence of upper body motion during speech. To make gesture units, we obtained two hours of motion captured data of a male and a female conversing. Instead of manually extracting gesture units, we defined set rules to extract the gesture units automatically. Rules for automatic extraction are: 1) Length is between 2 and 3 seconds. 2) starting and ending positions should have minimum distance. 3) should have optimal variance. Optimal variance is expert-dependent. The high variance will yield fewer but highly dynamic gesture units. The low variance will produce more but less dynamic units, including units with slight motion. Through this method, we extracted 2035 new gesture units with 1025 for male and 1010 for female styles. We observed that gestures could be clustered together to improve output diversity and keep relevance. We used Bisecting KMean algorithm to cluster the gesture units. Gesture clustering eliminated the TopK sampling of rules. Now sampling is done from the cluster.

2.2 2D Pose-Text pairs extraction

From public monologue-based videos with human presenters, we obtained the timestamped speech transcripts and 2D human pose with OpenPose[Cao et al. 2019]. Pose sequences were processed to make 3-second chunks (Text, Pose) pairs. The pose sequences are usually noisy and not usable directly. One way is to approximate grounding with predefined animations.

2.3 Rulemap generation with GestureCLR

We trained GestureCLR, a deep contrastive learning model based on the SimCLR[Chen et al. 2020] to match pose sequences with gesture units. For training, we used mocap data from the Talking with hands dataset[Lee et al. 2019]. The data was processed to make 2 to 3 seconds sequence units. Input to model was the clean 3D unit and 2D version of the same unit. The 2D version was further augmented with increasing gaussian noise and temporal shifts, thus mimicking the 2d pose data from the wild. The method's performance is compared with Mean Cosine Similarity in Figure 3. The model was applied to poses from (Text, Pose) pairs and gesture units from our mocap data to obtain (Text, Gesture) pairs. Text was converted to embedding with SentenceBERT[Reimers and Gurevych 2019] thus giving us (Text_{Emb}, Gesture).

3 IMPLEMENTATION

We used the server-client model to implement the gesture system. Unity is used as visualizing platform. Gesture units were converted from BVH to FBX and loaded in Unity animator. The gesture rule map is deployed on the server. The input text is chunked into 6-gram tokens and fed to the server's gesture system. The gesture system converts the token into embedding using sentencebert, finds the best match in the rulemap, and samples the gesture unit id from the corresponding cluster. The process is repeated for all the tokens, and ids are returned to the unity animator. Results can be seen here: https://youtu.be/QBtGdGE1Wgk

4 CONCLUSION

In this poster, we presented an elegant and easy-to-adopt system to generate co-speech gestures. We increased the gesture units from 54 to 2035 and rules from 84k to 210k from the baseline system. Since the rulemap is based on videos on various topics, it can easily be integrated into any domain possible. This method will improve the user experience while interacting with digital humans.

ACKNOWLEDGMENTS

This work was supported by the National Research Council of Science & Technology grant by the Korea government (NO. CRC-20-02-KIST) and also by the Korea Creative Content Agency in the Culture Technology Research & Development Program.

REFERENCES

Ghazanfar Ali, Hong-Quan Le, Junho Kim, Seung-Won Hwang, and Jae-In Hwang. 2019. Design of Seamless Multi-Modal Interaction Framework for Intelligent Virtual Agents in Wearable Mixed Reality Environment. In *CASA '19* (Paris, France). Association for Computing Machinery, New York, NY, USA, 47–52. https://doi.org/10.1145/3328756.3328758

Ghazanfar Ali, Myungho Lee, and Jae-In Hwang. 2020. Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds* 31, 4-5 (2020), e1944. https://doi.org/10.1002/cav.1944

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709* (2020).

Hanseob Kim, Ghazanfar Ali, and Jae-In Hwang. 2021. ASAP: Auto-generating Storyboard And Previz with Virtual Humans. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 316–320. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00071

Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 763–772. https://doi.org/10.1109/ICCV.2019.00085

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084