Real-Time Hand Gesture Recognition for OpenXR Using Transformer-Based Machine Learning

Salar Rezayani[1] Dr. Russell Butler[2]
[1]University of Waterloo [2]Bishop's University
srezayan@uwaterloo.ca · russell.butler@ubishops.ca

Abstract—We present a real-time hand gesture recognition system built on hand-tracking data captured through the OpenXR standard in Unity. Unlike video- or electromyography-based approaches, our method operates directly on 3D hand-joint positions and wrist rotation angles, which are embedded and processed by a Transformer encoder that models temporal dependencies within short frame windows. On a custom OpenXR dataset, the model reaches 93.7% gesture-classification accuracy, and normalizing joints relative to the palm makes recognition robust to hand orientation and size. We describe the architecture and data pipeline, define a protocol for benchmarking the method against public skeleton datasets, and outline how the approach can be extended toward detecting the flow of movement — the transitions between gestures — as future work.

***Index Terms*—**Hand gesture recognition, Transformers, OpenXR, virtual reality, human-computer interaction, machine learning.

I. INTRODUCTION

Gestures are an intuitive medium for human-computer interaction, and hand tracking is now widely available through the OpenXR standard. Classifying static gestures from joint positions is well studied, but detecting the flow of movement—the transitions and continuous motion between gestures—remains challenging. Transformers, which excel at modeling temporal dependencies, are a natural fit. This paper presents a Transformer-based system that recognizes hand gestures in real time, operating directly on OpenXR hand-joint data in Unity, and outlines how it can be extended toward flow-of-movement detection.

II. RELATED WORK

Transformers have been applied to hand gesture recognition across modalities: video and multimodal input (GestFormer [1]), surface electromyography (TraHGR [2]), and depth maps with surface normals [3]. Closest to our setting are skeleton-based methods that operate on 3D joints, such as DG-STA [4] and 3D-Jointsformer [5]. Our work differs by using OpenXR-tracked joints and wrist rotations captured in Unity for real-time recognition, with gesture-to-gesture flow detection identified as future work.

III. METHOD

A. Data Capture and Preprocessing

Gesture data is captured with Unity's OpenXR framework, recording the 3D positions of 21 hand joints and wrist rotation angles (pitch, yaw, roll). Joint positions are normalized relative to the palm and scaled by hand size, so recognition is invariant to hand dimensions and overall rotation. Sequences are segmented into overlapping windows of 30 frames to capture short-term motion context.

B. Transformer Architecture

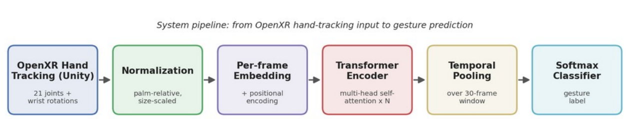

Fig. 1. System pipeline: from OpenXR hand-tracking input to gesture prediction.

Each frame's positional and rotational features are projected into a high-dimensional embedding, to which positional encodings are added. A stack of multi-head self-attention and feed-forward layers models dependencies across frames, and a softmax layer predicts the gesture label. Attention is computed as:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k})\, V,$$

where Q, K, and V are the query, key, and value matrices over frames and d_k is the key dimension. This lets the model focus on the frames most relevant to recognizing each gesture.

C. Training

The model is trained with the Adam optimizer and categorical cross-entropy loss on an 80/20 train/test split, using classification accuracy and F1-score over the gesture classes.

IV. RESULTS

The model reaches 93.7% gesture-classification accuracy, outperforming RNN and CNN baselines and reliably separating visually similar gestures such as "Point" and "Gun." Incorporating wrist orientation lets the system distinguish, e.g., an "Upward" from a "Neutral-Roll" pose, while palm-relative normalization keeps recognition stable across orientations; reliable orientation is, however, harder to obtain for some poses.

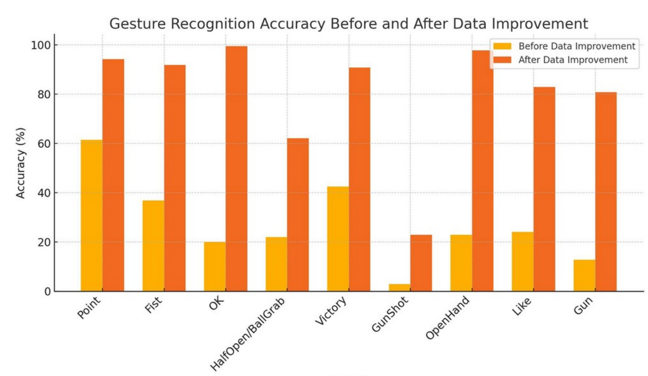

Fig. 2. Per-gesture accuracy before and after refining the data-collection process (coverage of rotations, sizes, and joint distances).

To position the method within the field and enable direct comparison, Table I lists published accuracies of representative skeleton-based methods on the SHREC'17 benchmark [6]. Evaluating our model under the same protocol is ongoing; the final row is a placeholder to be completed once those runs finish.

TABLE I. Reported accuracy (%) on SHREC'17 [6].

Method (venue)	14-cls	28-cls
ST-GCN (AAAI'18)	92.7	87.7

Method (venue)	14-cls	28-cls
Shift-GCN (CVPR'20)	95.5	89.4
CTR-GCN (ICCV'21)	96.1	94.4
TD-GCN (TMM'23)	97.02	95.36
DSTSA-GCN ('25)	97.74	95.37
Our method (this work)	—	—

V. CONCLUSION AND FUTURE WORK

We described a real-time, Transformer-based hand gesture recognition system that operates on OpenXR hand-tracking data. Palm-relative normalization and attention over time yield accurate, orientation-robust recognition. Future work includes low-latency deployment in Unity via ONNX/Barracuda, evaluation on public skeleton benchmarks (SHREC'17, DHG-14/28) to obtain directly comparable numbers, and extending the model from isolated-gesture classification toward flow-of-movement detection — the transitions between gestures — and continuous motions such as sign language.

DECLARATIONS

Ethics and consent. The hand-tracking data was collected by the authors and consists solely of hand-joint positions and wrist rotations captured in Unity via OpenXR; it contains no images, faces, or other personally identifiable information. As the data comprises only non-identifying hand-motion recordings gathered by the researchers, no separate ethics-board review was sought.

Use of AI tools. AI-based tools assisted with drafting and editing portions of this manuscript, including language, figures, and reference formatting. All technical content, system design, data collection, experiments, and results are the authors' own, and were reviewed and verified by the authors.

REFERENCES

[1] M. Garg, D. Ghosh, and P. M. Pradhan, "GestFormer: Multiscale Wavelet Pooling Transformer Network for Dynamic Hand Gesture Recognition," in Proc. CVPR Workshops, 2024, arXiv:2405.11180.

[2] S. Zabihi, E. Rahimian, A. Asif, and A. Mohammadi, "TraHGR: Transformer for Hand Gesture Recognition via Electromyography," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 31, pp. 4211–4224, 2023.

[3] A. D'Eusanio, A. Simoni, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara, "A Transformer-Based Network for Dynamic Hand Gesture Recognition," in Proc. Int. Conf. 3D Vision (3DV), 2020, pp. 623–632.

[4] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas, "Construct Dynamic Graphs for Hand Gesture Recognition via Spatial-Temporal Attention," in Proc. BMVC, 2019, arXiv:1907.08871.

[5] E. Zhong, C. R. del-Blanco, D. Berjón, F. Jaureguizar, and N. García, "Real-Time Monocular Skeleton-Based Hand Gesture Recognition Using 3D-Jointsformer," Sensors, vol. 23, no. 16, art. 7066, 2023.

[6] H. Cui, R. Huang, R. Zhang, and T. Hayama, "DSTSA-GCN: Advancing Skeleton-Based Gesture Recognition with Semantic-Aware Spatio-Temporal Topology Modeling," arXiv:2501.12086, 2025.