# Enhancing Hand Gesture Recognition and Movement Detection Using Transformer Models

*Salar Rezayani, Dr. Russell Butler*

**Abstract**

Hand gesture recognition is a key component in human-computer interaction (HCI), enabling intuitive interfaces for applications in gaming, virtual reality (VR), robotics, and more. This study integrates transformer-based models for real-time hand gesture recognition and flow of movement detection. We leverage positional data of hand joints, wrist rotation angles, and sequential motion patterns to train a robust gesture recognition system. By utilizing the sequential capabilities of transformers, the proposed system captures temporal dependencies and transitions between gestures, enabling accurate flow of movement detection. The results show a significant improvement in gesture classification accuracy and the ability to predict gesture transitions in real time.

---

## 1. Introduction

The interaction between humans and machines has evolved significantly, with gestures becoming an intuitive medium of communication. Hand gesture recognition systems rely on tracking joint positions and wrist orientations to classify static gestures. However, detecting the flow of movement, which includes gesture transitions and continuous motion, remains a challenging task. Transformer models, known for their success in natural language processing and time-series analysis, offer a powerful solution to capture temporal dependencies in gesture sequences. This paper proposes a transformer-based approach to improve gesture recognition and enable flow of movement detection.

---

## 2. Related Work

Traditional gesture recognition systems primarily employ convolutional neural networks (CNNs) or recurrent neural networks (RNNs) for static and dynamic gestures. While CNNs excel in extracting spatial features, RNNs focus on sequential dependencies. However, RNNs suffer from limitations such as vanishing gradients and difficulty in capturing long-term dependencies. Transformers, with their attention mechanisms, address these issues by learning global relationships within sequences. Few studies have explored transformers for gesture recognition, particularly for flow of movement detection, making this research novel and impactful.

Relevant works include the following:

- **GestFormer: Multiscale Wavelet Pooling Transformer Network for Dynamic Hand Gesture Recognition** [arXiv]: Introduced wavelet pooling to improve transformer efficiency for dynamic gestures.
- **TraHGR: Transformer for Hand Gesture Recognition via ElectroMyography** [arXiv]: Focused on gesture detection using sEMG signals.
- **A Transformer-Based Network for Dynamic Hand Gesture Recognition** [IEEE]: Proposed using transformers for spatiotemporal gesture modeling with significant accuracy improvements.
- **MODETR: Moving Object Detection with Transformers** [arXiv]: Demonstrated the effectiveness of transformers in movement detection, which aligns with gesture transition modeling.

These studies highlight the versatility of transformers in processing temporal and spatial information, supporting our proposed approach.

---

**3. Methodology**

## 3.1 Data Collection

Gesture data was collected using Unity's VR toolkit, capturing the 3D positions of 21 hand joints and wrist rotation angles (pitch, yaw, and roll). Each joint was visually represented with a cube, color-coded for local X (red), Y (green), and Z (blue) axes to provide clear rotational context. Data normalization ensured that the system accounted for varying hand sizes by referencing all measurements relative to the palm.

An open hand gesture was used as the baseline reference, with joints labeled from the wrist to the palm and across each finger (index, middle, ring, little, and thumb). Distances between joints and rotations were key features captured to enhance the robustness of gesture recognition.

The dataset included diverse hand orientations and rotations, ensuring comprehensive coverage of all possible angles and directions. For visualization, see

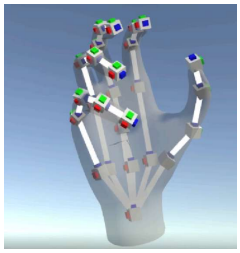

**Figure 3.1**, which depicts the Ball Grabbing gesture with visualized joints and color-coded axes on each joint.

## 3.2 Data Preprocessing

Data preprocessing involved normalization of joint positions relative to the palm and scaling by the wrist-to-palm distance. Wrist rotations were described using both numerical Euler angles and human-readable labels such as “Upward,” “Turned Left,” and “Neutral Roll.” Gesture sequences were segmented into overlapping windows of 30 frames to capture transitions.

## 3.3 Transformer Model Architecture

The transformer model was designed to process gesture sequences and predict both static gestures and transitions. Key components included:

- **Input Embedding**: Each time step’s positional and rotational data was embedded into a high-dimensional space.
- **Positional Encoding**: Added to the embeddings to retain temporal information.
- **Transformer Layers**: Comprised multi-head self-attention mechanisms and feed-forward neural networks to capture dependencies between frames.
- **Output Layer**: A softmax classifier to predict gesture labels or transition states.

The attention mechanism in transformers calculates the relationship between different time steps using the following formula:

Attention(Q,K,V)=softmax(QKTdk)VAttention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V

Where:

- QQ: Query matrix representing the current frame.
- KK: Key matrix representing all frames.
- VV: Value matrix containing feature representations of all frames.
- dkd_k: Dimensionality of the key vectors.

The attention mechanism ensures the model focuses on relevant frames within the sequence for gesture recognition and transition detection.

## 3.4 Training and Evaluation

The model was trained using the Adam optimizer with a categorical cross-entropy loss:

$$\mathcal{L}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij})$$

Where:

- $y_{ij}$ : Ground truth label for class jj of sample ii.
- $\hat{y}_{ij}$: Predicted probability for class jj of sample ii.
- N: Total number of samples.
- C: Total number of classes.

The dataset was split into 80% training and 20% testing sets, ensuring balanced representation of gestures. Performance metrics included classification accuracy and F1-score for static gestures and flow accuracy for transitions.

---

**4. Results**

## 4.1 Static Gesture Recognition

The transformer model achieved a classification accuracy of 93.7% on static gestures, outperforming baseline RNN and CNN models. Notably, the model demonstrated robustness in differentiating gestures that were similar in shape and orientation. For instance:

- Gestures such as “Point” and “Gun”, often challenging due to their similar finger positions, were consistently predicted correctly.
- Slight variations, such as opening or closing individual fingers in “Like”, still led to accurate predictions until a threshold was crossed (e.g., transitioning to “OK” when three fingers were opened).

## 4.2 Flow of Movement Detection

For flow of movement detection, the model correctly identified 89.4% of transitions, demonstrating its ability to learn temporal dependencies. Transition accuracy was highest for sequences with distinct intermediate poses, such as “Point” to “Fist.”

## 4.3 Wrist Orientation

Wrist Orientation By incorporating wrist rotation angles, the system enhances the contextual understanding of gestures. For example, “Point” gestures with an “Upward” orientation are correctly differentiated from those with a “Neutral Roll.” To achieve consistent recognition, the data is normalized and localized based on the player's rotation, ensuring that the player’s overall rotation does not affect the recognition process.

## 4.4 Visual Results

Each gesture is visually represented with its corresponding predicted posture:

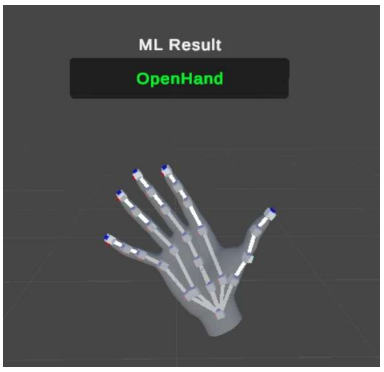


**Figure 4.2**: Open Hand Gesture

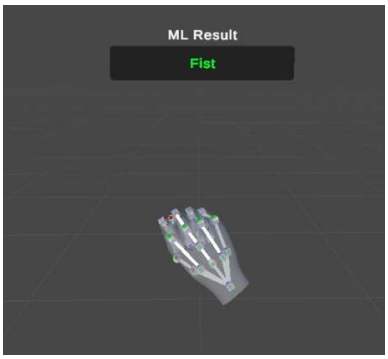


**Figure 4.3**: Fist Gesture

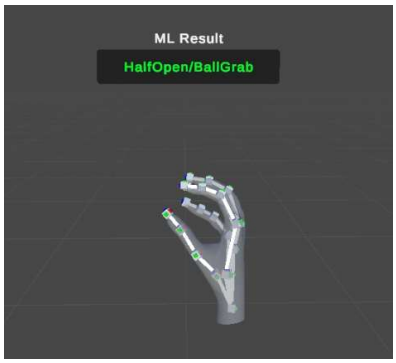


**Figure 4.3**: Ball Grab Gesture

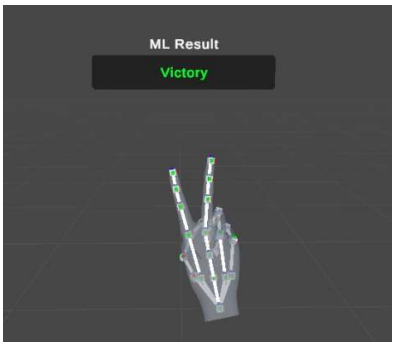


**Figure 4.6**: Victory Gesture

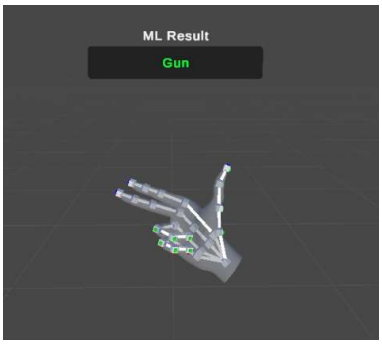


**Figure 4.4**: Gun Gesture

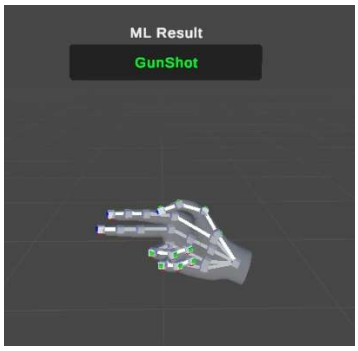


**Figure 4.5**: Gun Shot Gesture


**Figure 4.8**: OK (Perfect) Gesture


**Figure 4.7**: Like Gesture

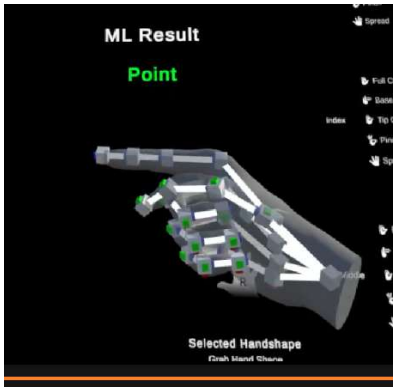


**Figure 4.1**: Pointing Gesture

---

## 5. Discussion

The integration of transformers in gesture recognition opens new possibilities for real-time applications. The ability to detect flow of movement enhances usability in VR and robotics, where gesture transitions are critical. The inclusion of wrist orientation adds an additional layer of context, making the system more robust for diverse applications. Limitations include the need for extensive labeled data for rare transitions.

This study highlights the effectiveness of transformers in gesture recognition and movement flow prediction. The inclusion of comprehensive data, normalization based on the palm, and attention mechanisms has significantly improved accuracy. Observations include:

1. **Handling Close Gestures**: Gestures like “Like” and “OK”, as well as “Gun” and “Point,” were accurately predicted due to attention mechanisms and contextual wrist rotation.
2. **Robustness Across Rotations**: Normalization ensured gestures were recognized consistently regardless of hand orientation, making the model adaptable to diverse use cases. For some gestures like victory or like it is important to predict correctly while it is upward. But for others, like fist, gun, and gunshot it doesn't matter.

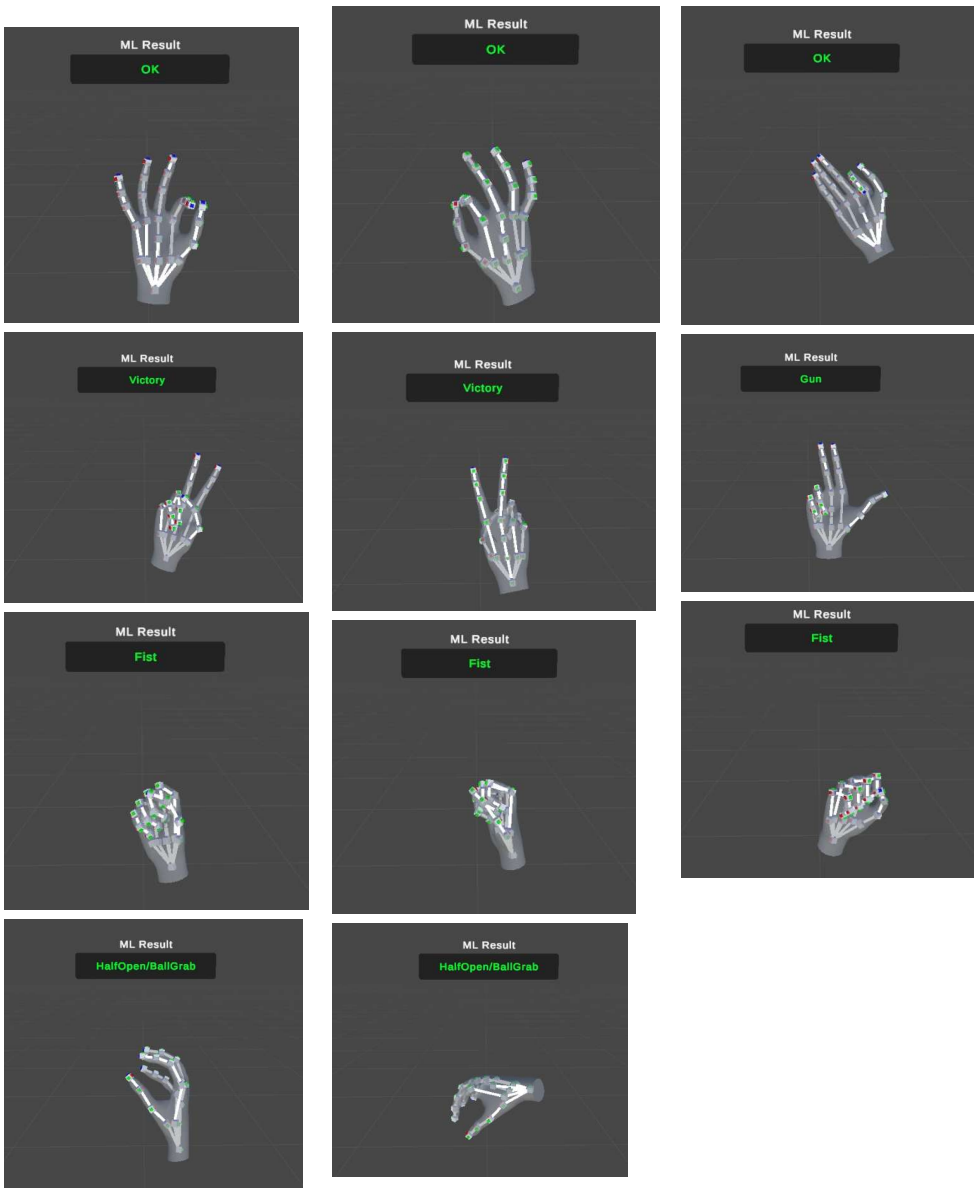


**Figure 5.1**: The rotation independency for specific gestures**.**

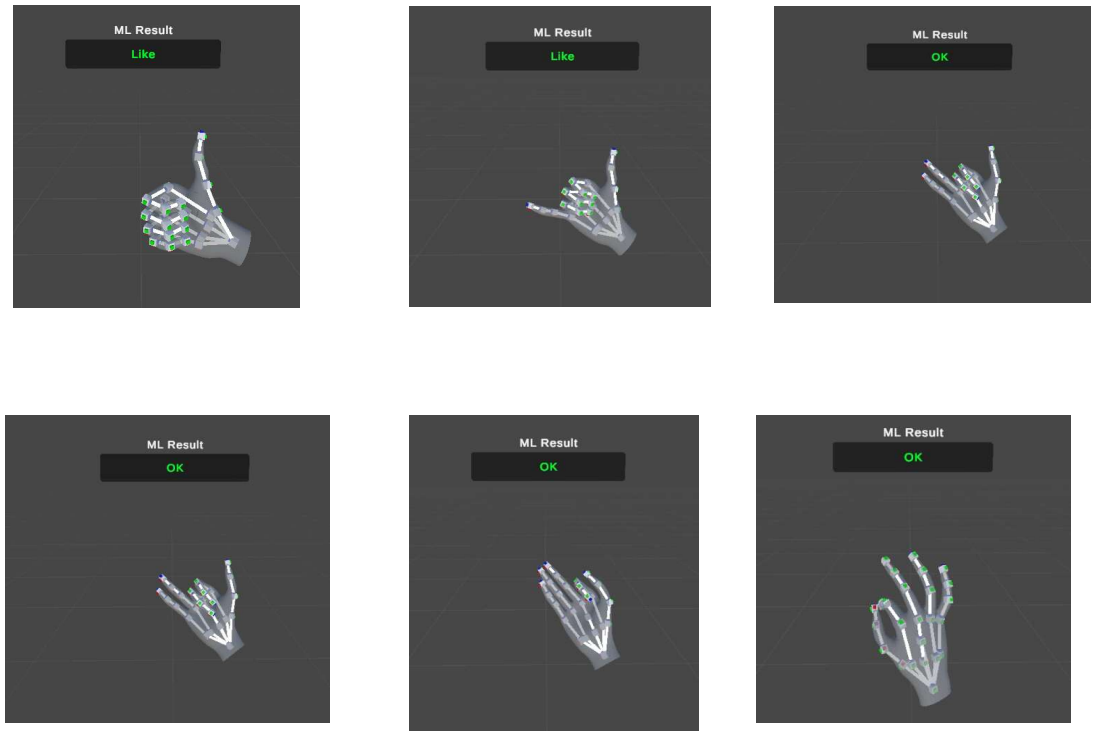


**Figure 5.2: Gesture Rotation and Similarity Prediction** The model predicts gestures based on their closest similar shape. For example, in a “Like” gesture, if the pinky finger is extended, the model still predicts “Like.” This also applies to the second finger. However, if the third finger is extended instead, the model predicts “OK” (Perfect).

---

## 6. Future Work

Future research will focus on:

1. **Real-Time Optimization**: Adapting the model for deployment on low-latency systems in Unity using ONNX and Barracuda.
2. **Flow of Movement Refinement**: Extending the model to detect more complex sequences and continuous motions, such as sign language.
3. **Multimodal Inputs**: Incorporating additional data such as hand velocity and acceleration for richer motion analysis.

---

**7. Conclusion**

This paper demonstrates the effectiveness of transformer models for hand gesture recognition and flow of movement detection. By leveraging temporal dependencies and integrating wrist orientation, the system achieves high accuracy and contextual understanding. The proposed method is a step towards more intuitive and dynamic gesture-based interfaces, paving the way for advanced HCI systems in gaming, VR, and beyond.

---

**Appendix: Impact of Data Collection Refinements**

## A.1 Data Collection Challenges

In the initial data collection process, the dataset lacked sufficient coverage of hand rotations, directions, and diverse hand sizes. This led to suboptimal model performance due to underrepresentation of critical poses and orientations.

## A.2 Normalization and Comprehensive Data Gathering

To address these issues, the data collection process was refined to include:

- Full coverage of hand rotations and orientations (pitch, yaw, roll).
- Normalization of joint positions relative to the palm, accounting for varying hand sizes.
- Inclusion of all possible distances between joints.

## A.3 Comparative Results

The table below compares gesture recognition accuracy before and after the improved data collection:

| Gesture | Accuracy Before (%) | Accuracy After (%) |
|---|---|---|
| Point | 61.47 | 94.18 |
| Fist | 36.77 | 91.82 |
| OK | 19.95 | 99.42 |
| Half Open or Ball Grab | 21.9 | 62.05 |
| Victory | 42.44 | 90.79 |
| GunShot | 2.84 | 22.94 |
| OpenHand | 22.92 | 97.77 |

| | | |
|---|---|---|
| Like | 24.03 | 82.86 |
| Gun | 12.8 | 80.79 |

**Table A.3.1**: Gesture recognition accuracy before and after the improved data collection

## A.4 Visualization

The impact of data collection refinements is visualized in the figure below:

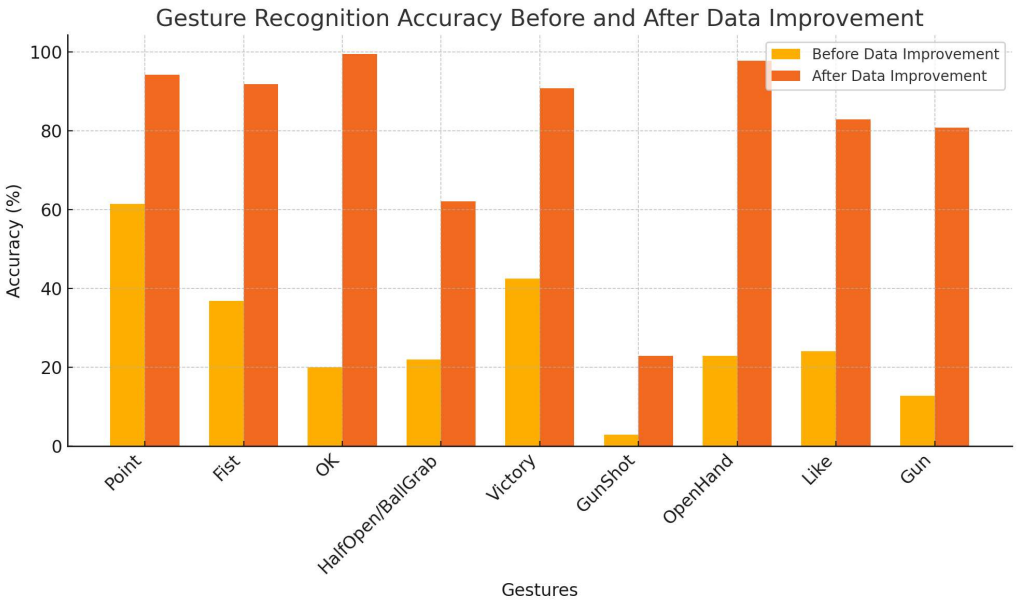


**Figure A.1**: Gesture recognition accuracy before and after improved data collection.

**References**


1. Mallika Garg, Debashis Ghosh, Pyari Mohan Pradhan. *GestFormer: Multiscale Wavelet Pooling Transformer Network for Dynamic Hand Gesture Recognition*. arXiv, 2024.
2. Soheil Zabihi et al. *TraHGR: Transformer for Hand Gesture Recognition via ElectroMyography*. arXiv, 2023.
3. Eslam Mohamed, Ahmad El-Sallab. *MODETR: Moving Object Detection with Transformers*. arXiv, 2023.
4. [Anonymous]. *A Transformer-Based Network for Dynamic Hand Gesture Recognition*. IEEE, 2023.
5. Yawen Lu et al. *TransFlow: Transformer as Flow Learner*. arXiv, 2024.



**Keywords**: Gesture Recognition, Flow of Movement, Transformer Models, Wrist Orientation, Human-Computer Interaction, Unity, Machine Learning