Leveraging Self-Attention Mechanism for Deep Learning in Hand-Gesture Recognition System

Muhamad Amirul Haq[1*], *Le Nam Quoc* Huy[2], *Muhammad* Ridlwan[1], and *Ishmatun* Naila[1]

[1]Universitas Muhammadiyah Surabaya, Surabaya, Indonesia
[2]National Taiwan University of Science and Technology, Taipei, Taiwan

Abstract. This research addresses the complex challenge of recognizing hand gestures irrespective of the user's body posture, a crucial issue in medical treatment for people with speech impairments and human-machine interfaces where precise gesture interpretation is vital. The aim is to engineer an advanced hand gesture recognition system, effective across various body positions and camera viewpoints. A novel flexible camera arrangement was employed, integrating a CNN-Transformer hybrid model, leveraging the strengths of Convolutional Neural Networks and the self-attention mechanism of Transformers. Developed using Python and the PyTorch deep learning framework, the system focuses on sophisticated image processing techniques. A thorough literature review on gesture recognition systems and multi-view analysis was conducted to inform the development. The system demonstrated exceptional accuracy in recognizing hand gestures in diverse body postures and from multiple camera perspectives, significantly outperforming existing methods. It marked a significant advancement in decoding complex gestures, a key aspect for medical applications and intricate human-machine interactions. This technology is primarily beneficial for people with speech impairments, rehabilitation, and in human-machine interfaces, poised to revolutionize patient care and enhance interaction with advanced machinery and computer systems.

1 Introduction

Most electronic device interfaces in current days rely on touch-based input methods such as physical buttons on remotes or touch panels on devices like smartphones [1]. This approach has become an industry standard; however, the primary challenge faced is the limitation in the number of controls that such interfaces can offer. This limitation is significantly influenced by the physical size of the control devices. This issue becomes particularly pronounced in educational environments, especially in schools, where interaction with electronic equipment forms a crucial part of the learning process [2]. When situations arise where multiple users, specifically students and teachers, wish to control electronic devices simultaneously, the need to use control devices in turns, or in some cases, to provide multiple control devices corresponding to the number of users, becomes evident. This not only hinders the efficiency of the learning process but also can lead to costly and short-term investments,

* Corresponding author: amirulhaq@ft.um-surabaya.ac.id

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

considering the rapid advancement in technology, especially those related to display technologies. To address these challenges, various alternative and universal electronic equipment control systems have been developed [3]. A prominent solution among these is the implementation of visual controls, which allow users, particularly students and teachers, to interact with electronic devices using hand gestures and other visual elements. More intuitive and user-friendly devices enhance the quality of learning by reducing distractions caused by complicated control tools.

Hand gesture tracking and recognition systems require models capable of processing both spatial and temporal data [4]–[7]. Spatial data consists of the location and pose of the hand, while temporal data involves the sequence of hand movements over time. Traditionally, CNNs have been the state-of-the-art method for processing spatial data [8], [9], whereas Long-Short Term Memory (LSTM) networks have excelled in processing temporal data [10], [11]. However, LSTM's performance has been exceeded by the self-attention mechanism in recent years. Therefore, using the self-attention mechanism to process the temporal data would be able to improve the accuracy of the algorithm.

The last decade's advancements in computer vision have opened significant opportunities to enhance electronic device interface methods in schools. In this context, the self-attention mechanism of Transformers holds great potential in providing a reliable and efficient solution [12]. Transformers are a type of neural network architecture that has been exceptionally successful in various applications, including natural language processing and image recognition. Utilizing the self-attention mechanism of Transformers in the context of hand gesture recognition and tracking has the potential to enhance the accuracy and reliability of the system, which is crucial in educational environments. Besides its potential in electronic device interfaces, the system designed by the researchers also has positive implications for education. By reducing physical contact with electronic devices, this system can contribute to enhancing the effectiveness of learning, particularly in an era where technology is an integral component of the educational process.

Overall, this research signifies a pivotal evolution in how schools and educational institutions interact with electronic devices. Moreover, this research can also be used to translate hand signs for hearing or speech-impaired people. By leveraging advancements in computer vision and the Transformer's self-attention technology, this system has the potential to improve the accuracy and performance of the existing algorithms. With a more efficient and user-friendly solution, this research is expected to make a tangible contribution to enhancing the quality of learning in educational and medical environments.

2 Method

2.1 Overview

This study focuses on developing a hand gesture recognition system using transformers, capable of translating a sequence of individual hand-signs into meaningful commands or actions. The system is designed to work in a Human-Machine Interface (HMI) setup, leveraging monocular camera inputs for comprehensive visual data acquisition. The overall architecture integrates advanced technologies such as the Mediapipe Hands Framework and Transformer Sequence Translator, as illustrated in **Fig. 1**.

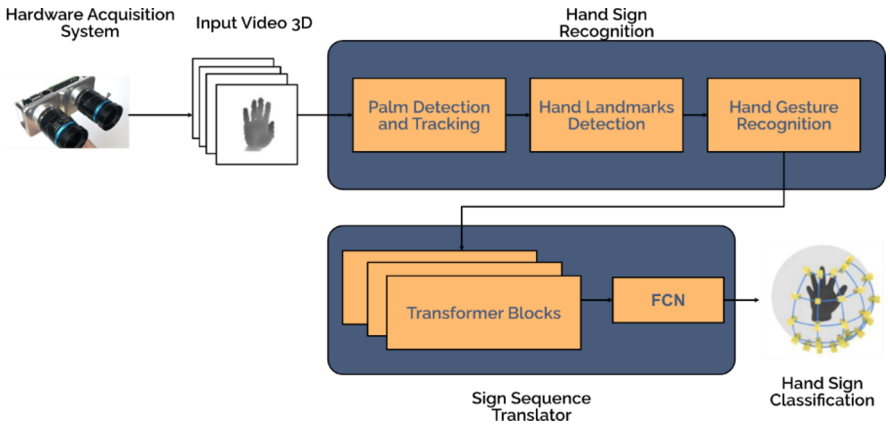

Fig. 1 Overview of network architecture

2.2 Hand gesture recognition

The hand gesture recognition component is a crucial part of the system, built on the Mediapipe Hands Framework. This framework provides a robust foundation for detecting hand poses with high accuracy. It consists of two main stages: palm detection, hand landmarks detection, and hand gesture recognition. The palm detection phase employs an auto-encoder network, it provides fast identification of the presence and orientation of the hand. Following this, the hand landmarks detection and gesture recognition stages utilize a ResNet-based feature extractor that is shared with the hand gesture recognition part. This feature extractor is trained to recognize specific hand gestures, including "Thumbs Up," "Victory," and "Open Palm." These gestures are chosen for their distinctiveness and relevance in common communication.

2.3 Transformer for sequence translation

The transformer model utilized in this system is based on the self-attention mechanism, which is central to its ability to process sequences of hand gestures. We employ a lightweight transformer model that consists of three encoder and decoder layers. Then, the fully connected layer by the end of the network maps the output into 39 classes, each corresponding to a different command. The self-attention mechanism in the transformer is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In this formula, Q, K, and V represent the query, key, and value matrices, respectively, derived from the input sequence. The softmax function applies to the scaled dot product of Q and K^T, divided by the square root of the dimension of the keys, d_k. This scaling is crucial for stabilizing the gradients during the learning process. The resulting output is then multiplied by the value matrix V, producing an output that captures the contextual relationships within the input sequence.

2.4 Network's training and testing

The training of the Transformer model employs the Cross-Entropy Loss function, which is particularly suitable for classification tasks. The Cross-Entropy Loss is defined as:

$$\text{Loss} = -\sum_{c=1}^{M} y_{o,c} \log\left(p_{o,c}\right) \tag{2}$$

Here, M represents the number of classes. In this implementation, there are three individual hand signs and 39 unique sequences that can be obtained from the combination of the individual hand gesture. For each observation o and class c, $y_{o,c}$ is a binary indicator (0 or 1) that specifies whether class c is the correct classification for observation o. The term $p_{o,c}$ is the predicted probability that observation o belongs to class c. The loss is computed as the negative sum of the logarithm of the predicted probabilities, specifically for the true class labels. This formulation of the loss function effectively drives the optimization process by penalizing incorrect classifications and reinforcing correct ones. The loss for every iteration has been recorded and shown in **Fig. 2**. It shows that the model can learn effectively as the loss keeps going down over time.

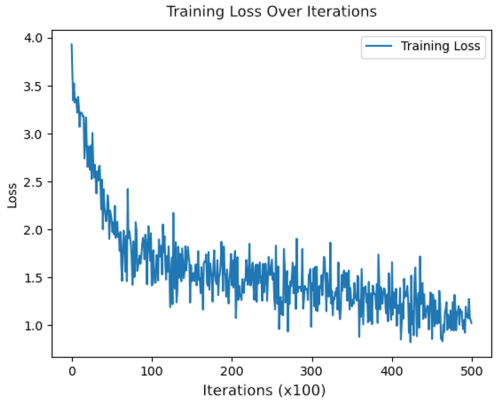

Fig. 2. Training loss over iterations

3 Result and discussion

3.1 Experimental setup

The proposed method is trained on a Windows 11 desktop with Intel i7 13700KF CPU, 64 GB RAM, and NVidia RTX 3060-12GB VRAM GPU. The proposed methods are written using Pytorch [13] framework for the Sequence Translation's Transformer part and with Google's MediaPipe [14] for gesture recognition. The proposed method is trained for 50,000 iterations with an AdamW optimizer. Based on the empirical results, the best initial learning rate to train our model is 10^{-4}.

3.2 Discussions

Our initial testing, as detailed in the accompanying **Table 1-3**, has shown promising results for our method's ability to recognize hand gestures, achieving an accuracy of up to 75%. Notably, the algorithm's lightweight design enables it to operate in real-time, maintaining a

reasonable real-time inference at 28 frames per second (FPS), which is crucial for seamless user interaction. However, the accuracy of the sign sequence translator currently stands at 64.8%, indicating room for improvement. Moreover, the model struggles to recognize "Thumbs Up" as it has multiple fingers intertwined and obstruct each other. This condition causes the model to have difficulty recognizing the hand landmark and classifying the gesture. Using higher-quality cameras or standing closer to the lens has been observed to mitigate this problem. The visualized results of our model prediction is shown in **Fig. 3.**

Table 1. Accuracy of each gesture

	Gesture 1	**Gesture 2**	**Gesture 3**
Accuracy	25%	100%	100%

Table 2. Accuracy of each subject

	Subject 1	**Subject 2**	**Subject 3**	**Subject 4**
Accuracy	66.7%	66.7%	66.7%	100%

Table 3. Summary of the model's performance

Item		**Specification**
Desktop Hardware	:	Intel i7-13700KF, Nvidia RTX 3060 12GB
Camera	:	Logitech Brio 500
Input Dimension	:	1920 × 1080
FLOPs	:	70 MB/s
Individual hand-sign accuracy (3 classes)	:	75%
Hand-sign sequence accuracy (39 classes)	:	64.8%
Average runtime	:	28 FPS

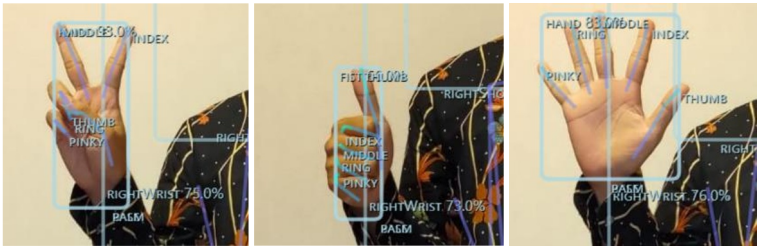

Fig. 3. The system can recognize 39 unique sequences, each corresponding to a different command, by combining three or fewer distinct hand gestures.

4 Conclusion

In conclusion, this research provides an alternative modality in the field of human-machine interface, specifically within educational environments. Our proposed method, leveraging the self-attention mechanism of the Transformer model, has demonstrated its potential in revolutionizing the way electronic devices are interfaced in schools and educational institutions. The initial results, as evidenced by our limited testing, underscore the feasibility and effectiveness of our approach. With an accuracy of up to 75% in hand gesture recognition and the capability to operate at a real-time rate of 28 FPS, the system shows considerable promise in facilitating seamless and intuitive interactions. The Sign Sequence Translator, despite its current accuracy of 64.8%, represents a substantial advancement in interpreting sequences of hand gestures. The transformative potential of this technology lies in its ability

to bridge the gap between complex human gestures and digital commands, thereby creating a more natural and accessible mode of interaction with technology.

Looking ahead, the system's capabilities can be significantly enhanced through more rigorous testing and training with an expanded set of gestures. By doing so, the range of commands and the breadth of the sign language vocabulary that the system can recognize and interpret will be considerably broadened. This enhancement will not only elevate the efficiency of the learning process but also contribute to a more inclusive and technologically integrated educational environment.

References

1. A. A. Bello, H. Chiroma, A. Y. Gital, L. A. Gabralla, S. M. Abdulhamid, and L. Shuib, "Machine learning algorithms for improving security on touch screen devices: a survey, challenges and new perspectives," *Neural Comput. Appl.*, vol. 32, no. 17, pp. 13651–13678, Sep. 2020, doi: 10.1007/s00521-020-04775-0.
2. "Sensors | Free Full-Text | Review of Capacitive Touchscreen Technologies: Overview, Research Trends, and Machine Learning Approaches." Accessed: Oct. 31, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4776
3. [W. Ding, C. Wang, B. Fang, F. Sun, and J. Shan, "A Survey of Multimodal Human-Machine Interface," in *Cognitive Systems and Signal Processing*, F. Sun, H. Liu, and B. Fang, Eds., in Communications in Computer and Information Science. Singapore: Springer, 2021, pp. 379–386. doi: 10.1007/978-981-16-2336-3_35.
4. L. Ge *et al.*, "3D Hand Shape and Pose Estimation From a Single RGB Image," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10833–10842. Accessed: Oct. 31, 2023. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Ge_3D_Hand_Shape_and_Pose_Estimation_From_a_Single_RGB_CVPR_2019_paper.html
5. C. Zheng *et al.*, "Deep Learning-based Human Pose Estimation: A Survey," *ACM Comput. Surv.*, vol. 56, no. 1, p. 11:1-11:37, Aug. 2023, doi: 10.1145/3603618.
6. X. Chen, G. Wang, H. Guo, and C. Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *Neurocomputing*, vol. 395, pp. 138–149, Jun. 2020, doi: 10.1016/j.neucom.2018.06.097.
7. C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RGB Images," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 813–822. Accessed: Oct. 31, 2023. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Zimmermann_FreiHAND_A_Dataset_for_Markerless_Capture_of_Hand_Pose_and_ICCV_2019_paper.html
8. "ImageNet Classification with Deep Convolutional Neural Networks (AlexNet) - actorsfit." Accessed: Oct. 31, 2023. [Online]. Available: https://actorsfit.com/a?ID=00450-9b1db208-6e87-450e-bc1d-78f9c16d5996
9. J. Lin, C. Gan, and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7083–7093. Accessed: Oct. 31, 2023. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Lin_TSM_Temporal_Shift_Module_for_Efficient_Video_Understanding_ICCV_2019_paper.html

10. Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019, doi: 10.1162/neco_a_01199.
11. B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, pp. 650–655, Jan. 2021, doi: 10.1016/j.procir.2021.03.088.
12. A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Oct. 31, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
13. A. Paszke et al., "Automatic differentiation in PyTorch," Oct. 2017, Accessed: Jan. 02, 2024. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ
14. C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," arXiv.org. Accessed: Jan. 02, 2024. [Online]. Available: https://arxiv.org/abs/1906.08172v1