A Novel Approach for Virtual Locomotion Gesture Classification: Self-Teaching Vision Transformer for a Carpet-Type Tactile Sensor

Sung-Ha Lee* Ho-Taek Joo† Insik Chung Donghyeok Park Yunho Choi Kyung-Joong Kim

Gwangju Institute of Science and Technology (GIST)

ABSTRACT

Locomotion gesture classification in virtual reality (VR) is the process of analyzing and identifying specific user movements in the real world to navigate virtual environments. However, existing methods often necessitate the use of wearable sensors, which present limitations. To address this, we utilize a high-resolution carpet-type tactile sensor as a foot action recognition interface, which is previously unexplored in the context of locomotion gesture classification. This interface can capture the user's foot pressure data in detail to distinguish similar actions. In this paper, to efficiently process captured user's foot tactile data and classify nuanced actions, we propose a vision transformer (ViT) architecture and a novel self-teaching vision transformer (STViT) model integrating elements of the shifted window vision transformer (SwinViT) and data-efficient image transformer (DeiT).

Index Terms: Human-centered computing—Human–computer interaction (HCI)—Interaction paradigms—Virtual reality

1 INTRODUCTION

Virtual locomotion is a fundamental component that significantly contributes to the immersive and interactive nature of the VR environment. Various studies have focused on exploring this field. However, despite the extensive research on locomotion methods, natural virtual locomotion continues to be a challenging domain. In an attempt to address the space constraints of limited real-world space, various in-place locomotion methods have been introduced. These methods involve users performing in-place gestures and translating these physical movements into a virtual world. However, in the field of classifying in-place gestures, the user is often required to wear sensors. The use of such wearable sensors can lead to discomfort as they disturb smooth movement.

Therefore, to recognize the user's actions without the need for wearable equipment, we utilized the carpet-type tactile pressure sensor developed by Luo *et al.* [1] as a foot action recognition interface. Its capacity to capture detailed foot pressure data holds the potential for discerning even subtle differences in similar user actions.

We utilized a vision transformer for processing tactile sensing data. To effectively classify in-place locomotion actions using tactile sensors, we propose the self-teaching vision transformer model (STViT), which is a novel architecture designed to utilize the benefits of both the DeiT and SwinViT.Our model incorporates a unique self-distillation technique that promotes the incremental accumulation of knowledge, leading to a stable learning process. This is achieved by using the training model itself from N-steps prior as the 'teacher' model, instead of using the separate pre-trained model as the teacher. Another noteworthy characteristic of our proposed model is the incorporation of patch merging, inspired by SwinViT, to enhance the ability to distinguish similar actions. This approach allows the model to change the size and focus of the areas it pays the most attention to, which helps it to better recognize variations in similar actions by capturing information at different scales. To evaluate performance, we compared three different algorithms which our self-teaching transformer model demonstrated the highest accuracy. The contributions of this work are as follows:

- We propose the self-teaching vision transformer model, which is a novel architecture that integrates the benefits of the DeiT and SwinViT. This model features a novel self-distillation technique that promotes incremental knowledge accumulation, leading to a more stable learning process.
- We utilized a carpet-type tactile sensor interface for foot action recognition, a novel application in the field of locomotion gesture classification. This high-resolution sensor captures detailed user foot pressure data, facilitating the differentiation of similar actions without the need to attach a sensor to the user's body.
- Our proposed model was evaluated against three different algorithms, demonstrating the best classification accuracy among them at 92.7%.

2 METHOD

2.1 Interface for Gesture Recognition

For the gesture recognition interface, we propose the use of carpet-type tactile pressure sensors developed by Luo *et al.* [1] to acquire the foot pressure data of users. This is the first time the carpet-type tactile sensor interface is used for VR locomotion gesture recognition. The selection of this sensor was motivated by its capacity to enhance user comfort as it doesn't necessitate wearing any sensor devices other than an HMD. Moreover, the high spatial resolution provided by the tactile sensor, which captures the human footprints in detail, allows for the recognition of similar locomotion actions.

2.2 Self-Teaching Vision Transformer (STViT)

As shown in Figure 1, we propose a novel approach that leverages the distillation technique to bolster the capability of detecting intricate patterns throughout various stages of SwinViT. The proposed method incorporates distilling the attention block and hidden state from a teacher model into a student model within the SwinViT framework. This suggests that the distillation of the attention block and hidden state from the teacher model to the student model affords a more nuanced update of the patch merging process. In our method, the teacher model is the student model from N-steps prior. Our method entails a more nuanced execution of patch merging procedures across several stages of SwinViT, thereby markedly boosting the model's representational capacity. This strategy effectively navigates the student model's learning process by harnessing the knowledge contained within the teacher model as a previous iteration of the student model.

Distillation of Self-Attention on Transformer: Drawing inspiration from [2], we chose to apply cross-entropy (CE) losses on the relationships among queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) in the MSA mechanism. More specifically, our initial step involves the

*e-mail: shlee0414@gm.gist.ac.kr

†e-mail: hotaek87@gm.gist.ac.kr

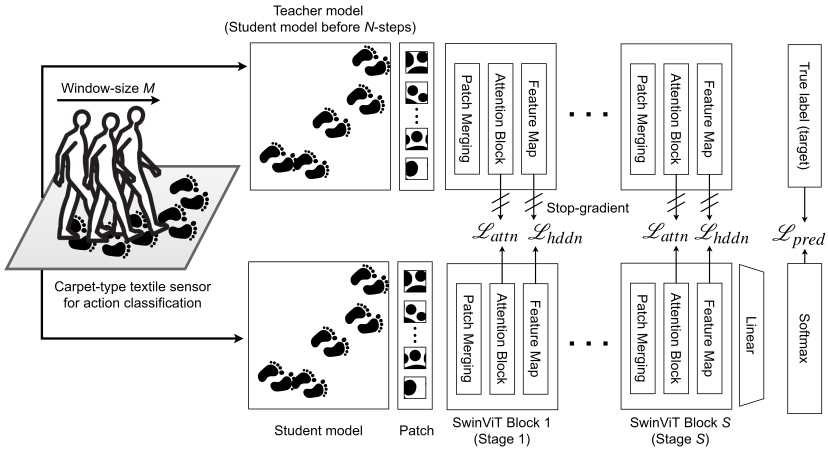

Figure 1: Architecture of our self-teaching vision transformer (STViT).

concatenation of matrices across all heads. For instance, we define $\mathbf{Q} = [\mathbf{Q}_1, \ldots, \mathbf{Q}_k] \in \mathbb{R}^{E \times kD}$, and $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{E \times kD}$ are defined in a similar manner. To simplify the notation, we employ $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ to denote $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively. Following this, we generate nine distinct relation matrices defined as $\mathbf{R}_{ij} = \text{softmax}(\mathbf{C}_i \mathbf{C}_j^T / \sqrt{kD})$. The self-attention distillation loss can be articulated as follows:

$$\mathscr{L}_{attn} = \frac{1}{9E} \sum_{n=1}^{E} \sum_{\substack{i \in \\ \{1,2,3\}}} \sum_{\substack{j \in \\ \{1,2,3\}}} CE(\mathbf{R}^s_{ij,n}, \mathbf{R}^t_{ij,n}) \tag{1}$$

where $\mathbf{R}^s_{ij,n}$ and $\mathbf{R}^t_{ij,n}$ represent the n^{th} rows of $\mathbf{R}_{ij}$ in the student and teacher models, respectively.

Distillation of Hidden States In this research, we define hidden states as each value between SwinViT Blocks. More specifically, the feature maps between these SwinViT stages are referred to as hidden states. The proposed method involves a process of distilling these hidden states from the teacher model $\mathbf{H}^t$ to the student model $\mathbf{H}^s$. At the end of each S stage, we utilize the CE loss function as defined by the following equation:

$$\mathscr{L}_{hddn} = \frac{1}{E} \sum_{n=1}^{E} CE(\mathbf{H}^s_n, \mathbf{H}^t_n) \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^{E \times D}$.

Prediction of Label-Logits The final component of the loss function is the CE loss between the true label and the prediction. This is effectively the loss incurred when images of window size M are input into the STViT model, comparing the difference between the model's prediction and the true label of the data.

$$\mathscr{L}_{pred} = CE(\mathbf{z}_s, \mathbf{y}) \tag{3}$$

In this equation, $\mathbf{z}_s$ and $\mathbf{y}$ represent the logits inferred by the student model and true label (target), respectively. Finally, we add the CE loss between the predicted and true labels. Consequently, we calculate the final objective function as

$$\mathscr{L}_{train} = \alpha \mathscr{L}_{attn} \times S + \beta \mathscr{L}_{hddn} \times S + \gamma \mathscr{L}_{pred} \tag{4}$$

In this formulation, the term S denotes the number of stages, whereas α, β, and γ are representative of hyperparameters, all of which are assigned a default value of 1.

2.3 Data Collection

We collected foot pressure data from 10 participants using carpet-type pressure sensors. The participants were asked to perform 7 tasks: walking, sneaking, marching, running, sliding, crouching, and standing. Each task was performed for two minutes on the carpet-type pressure sensor.

3 Model Evaluation

Table 1: Accuracy comparison of different models (%)

Model	Model Type	Mean	Standard Deviation
ResNet18	CNN	73.1	2.1
DeiT	ViT	89.4	3.2
SwinViT	ViT	89.8	1.3
STViT (Ours)	ViT	**92.7**	1.1

In this section, we present a comparative quantitative evaluation of various vision-based AI models, ranking them based on their accuracy, from highest to lowest in Table 1. Each model was trained for 40 epochs, and the accuracy is averaged over five runs.

4 Conclusions and Future Works

In this paper, we presented a novel method for VR locomotion gesture recognition, leveraging a carpet-type tactile sensor as an interface and proposing a novel STViT algorithm. This interface eliminates the need for wearable equipment, thereby enhancing user comfort and providing a more natural virtual experience. To process tactile data efficiently, we utilized a Vision Transformer and introduced the STViT algorithm. Our model integrates the advantages of both DeiT and SwinViT architectures and incorporates a self-distillation technique. Our model demonstrated improved performance in classifying locomotion actions, as validated by comparisons with three different algorithms.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) funded by the MSIT (2021R1A4A1030075) and the GIST-MIT Research Collaboration grant funded by the GIST in 2023.

References

[1] Y. Luo, Y. Li, M. Foshey, W. Shou, P. Sharma, T. Palacios, A. Torralba, and W. Matusik. Intelligent carpet: Inferring 3d human pose from tactile signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11255–11265, 2021.

[2] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2140–2151, 2021.