IMU Based Human Gestures Recognition using Deep Learning for Wearable Devices

Lin Fu
Victoria University of Wellington, Kelburn, New Zealand
fulin@myvuw.ac.nz

ABSTRACT

In recent years, the emergence of wearable devices, such as smartwatches and smart glasses. Based on this, human gesture recognition (HGR) for wearable devices has been a popular topic in the community of computer vision. Previous human gesture recognition approaches are mainly based on statistical methods, the breakthrough of deep learning enables researchers to make use of different neural network architectures to learn features of the human gestures that are collected from wearable devices. Most of these methods are based on the convolutional neural network, which extracts deep features from local manna. However, this way impedes the models from learning global information. To solve the problem, we propose a novel approach that leverages the attention mechanism to capture global information about human gestures. We conduct experiments on a benchmark, and the experimental results demonstrate that our proposed method is superior to the other baselines.

CCS CONCEPTS

• **Human-centered computing** → Ubiquitous and mobile computing; Ubiquitous and mobile computing design and evaluation methods.

KEYWORDS

deep learning, wearable devices, gestures, transformers

ACM Reference Format:
Lin Fu. 2022. IMU Based Human Gestures Recognition using Deep Learning for Wearable Devices. In *2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI 2022), December 16–18, 2022, Dongguan, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3584376.3584588

1 INTRODUCTION

In recent years, with the development of the Internet of Things, intelligent sensing technology has been widely used in wearable devices, which has promoted the development of wearable devices such as smartwatches. Since the introduction of Google Glass in 2012, various types of wearable devices have emerged, with richer functions and cheaper prices, and the number of shipments has been growing rapidly. According to the latest from Gartner, the global shipments of wearable devices were 141 million units in 2017 and 179 million units in 2018. The number of shipments is 179 million units in 2017 and will reach 225 million units in 2019. The number of smartwatch shipments has been in the first place. The rapid growth of the shipments of wearable devices, especially smartwatches, is in line with the actual application demand. The rapid growth of wearable devices, especially smart watches, is closely related to their actual application needs, including sports monitoring, health monitoring, location tracking, information notification, authentication, and assistance. The rapid growth in shipments of wearable devices, especially smart watches, is closely related to their practical application needs, including sports monitoring, health monitoring, location tracking, information notification, authentication and authentication, and auxiliary control.

Throughout the new interaction methods, gestures, eye movements, EEG, and body language have their own advantages and disadvantages. Among them, the hand, as the most flexible limb in the human body, can express rich meanings through gestures [1]. Gestures interaction can give users a higher sense of immersion while ensuring that the interaction is effective. HoloLens developed by Microsoft performs gesture recognition by image using the built-in camera to meet the human-to-human and human-to-nature interaction in various scenarios. This is also a typical application of gesture as a new human-computer interaction method. However, the HoloLens camera needs to be fixed in the head position and needs to meet a certain lighting environment, which will limit the scope of gesture interaction and application scenarios. In the digital society, users need a more convenient and natural interaction terminal without interference. In recent years, sensor-based wearable devices can be used in more scenarios without being affected by light or obstacles.

For traditional gesture recognition algorithms such as KNN [2] and SVM [3], they usually assign specific instructions to specific gestures, needing to determine the active gesture segment through the active segment detection mechanism, and analyse the data set in time and frequency domains with the help of complex feature extraction engineering, and finally achieving the correspondence between instructions and gestures through machine learning algorithm classification, which improves the accuracy of gesture recognition at the same time. However, when some changes occur in the data features, the accuracy of gesture recognition will be greatly reduced. The data collected by the independently designed wearable devices cannot cover all people.

The aforementioned recent research mainly rely on the convolution neural network to locally extract features of human gestures. The local feature extraction method ignores the global information, which is a crucial factor for recognising human gestures. Therefore, to address the severe problem, we make use of the transformers

RICAI 2022, December 16–18, 2022, Dongguan, China

ACM ISBN 978-1-4503-9834-3/22/12…$15.00
https://doi.org/10.1145/3584376.3584588

to further extract global information from human gestures and propose a novel fusion mechanism to fuse both the local and the global information, which leads to a better performance of the task.

2 BACKGROUND

2.1 Human Gesture Recognition

With the development of technology, as a pattern recognition problem, gesture recognition requires analysing the input data and learning valid feature data to recognise gesture postures. The current research on gesture recognition can be divided mainly into image data-based gesture recognition research and combined sensor-based gesture recognition research, depending on the type of data. [4].

The research on gesture recognition based on image data is mainly based on obtaining the user's gesture depth images and RGB images with the help of depth cameras, binocular cameras, RGB cameras, etc., to obtain the user's gesture depth images and RGB images and extracting the gesture image features based on the regional segmentation of the image. Extracting gesture image features, and using the learned features to process them by machine learning or deep learning to get the final. The learned features are processed by machine learning or deep learning to get the final recognition tag.

In recent years, many efforts have been done in the field of human gesture recognition. Xu et al. [5] proposed a gesture recognition algorithm based on RGB images and depth images. The hand was extracted from the image with the help of depth data and skin color features. The arm area in the image was removed by distance transformation to obtain the gesture picture of the combined palm and fingers, and the gesture data was recognised by the SVM algorithm. However, it can only achieve the recognition of static gestures.

The research on gesture recognition based on combined sensors mainly collects gesture motion data with the help of sensors and then implements gesture tag recognition with the help of machine learning algorithms through signal processing and feature extraction engineering. After processing and feature extraction engineering, gesture tag recognition is achieved by machine learning algorithms. The experience is not as natural as visual image-based gesture recognition based on visual images, but it can overcome the influence of light and occlusion, the device is wearable, and the recognition accuracy is also high. The current mainstream solutions include an inertial measurement unit (IMU), WIFI, and an arm ring. For instance, Fang et al. [6] et al. proposed a novel method for gesture capture and recognition based on inertial and magnetic measurement units. Through 18 low-cost IMMU combinations to capture and recognise the 3D movements of the arm, palm, and fingers, and the extreme learning machine (ELM), which is similar to the Broad Learning System. However, the system design is too complex for the whole arm. More recently, Utsumi A et al [7] used skeletal images taken by multiple cameras to hand position, gesture, and curvature of the hand were detected and proposed a method to track the 3D position, gesture, and shape of the hand from multi-viewpoint images. They proposed a method to track the 3D position, gesture, and shape of the hand from multi-viewpoint images, and the accuracy of detection was over 87.3%.

2.2 Transformers for Computer Vision

The transformers have been widely employed in the community of computer vision [8]. The transformers take the attention mechanism as the basis and consist of an encoder and a decoder. Attention is a mechanism in neural networks where models can learn to make predictions by selectively focusing on a specific set of data. Various attention mechanisms is proposed in recent years. The amount of attention is quantified by the learned weights, so the output usually forms a weighted average. In this paper, we mainly make use of self-attention. The self-attention is a kind of attention mechanism where the neural models leverage other observed parts of the same sample to make predictions about one part of the data sample. Conceptually, it feels very similar to non-local means. More importantly, the self-attention mechanism is variational and can be seen as an operation on sets. The equations are an exception to the prescribed specifications of this template. The calculation of the attention can be described as follows:

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

$$a_{ij}=\text{softmax}\left(\frac{\mathbf{q}_i\mathbf{k}_j^{\top}}{\sqrt{d_k}}\right) \tag{2}$$

The position embeddings of the transformer can be calculated as follows:

$$\text{PE}\left(i,\delta\right)=\begin{cases}\sin\left(\frac{i}{10000^{28}/d}\right) & \text{if}\quad \delta=2\delta'\\ \cos\left(\frac{i}{10000^{28}/d}\right) & \text{if}\quad \delta=2\delta'+1\end{cases} \tag{3}$$

3 PROPOSED METHOD

With the development of wearable devices, human-computer interaction technology has also evolved. Driven by the Internet of Things and artificial intelligence, gesture-based human-computer interaction technologies are gradually gaining popularity. New gesture-based interaction methods are constantly being researched both in academia and industry. With gesture interaction, people can interact with various devices more conveniently in different scenarios, including smart homes, smart vehicles, and virtual reality, which will greatly improve people's interaction experience. The existing gesture recognition methods can be classified into four categories based on the type of signals they use.

3.1 Preprocessing Module

Since the user causes a dramatic change in acceleration when making a gesture, by using a threshold-based approach, it is possible to find the start and end time points of a user's gesture in the dataset. Moreover, when extracting the acceleration, and angular velocity data during this time. Since the time used for handwriting different letters is different, resulting in the extraction of each letter corresponding to the amount of inertial sensing data extracted for each letter is not fixed, so it is necessary to exploit interpolation or cropping methods to ensure that the amount of inertial sensing data corresponding to all letters is the amount of inertial sensing data is fixed for all letters.

Besides, since the time used for handwriting different letters is different, the amount of inertial sensing data generated is different.

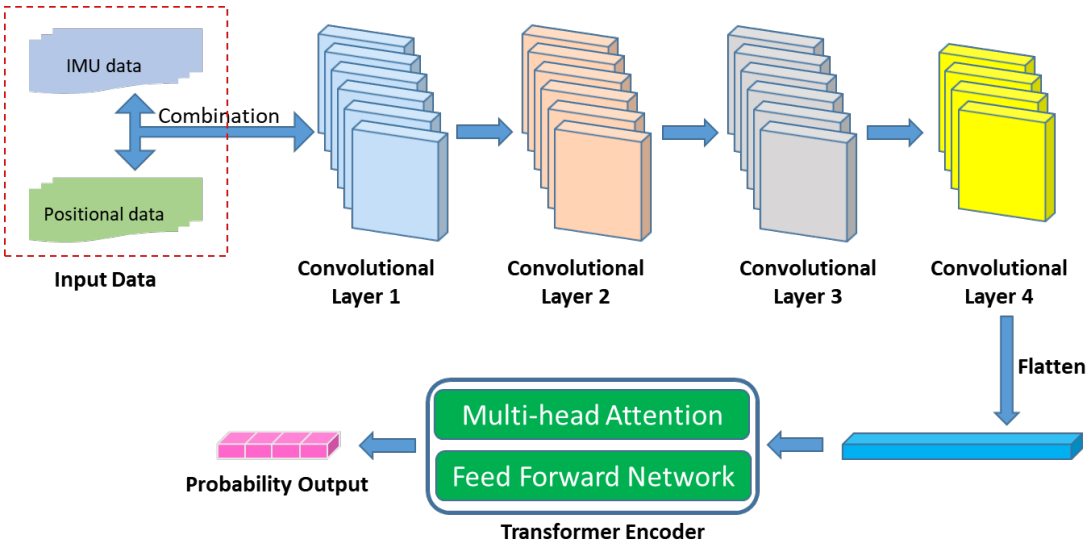

Figure 1: The Architecture of The Proposed Method TranC

The size of array A obtained by data slicing is not fixed, and the size of the inertial sensing data input to our model is fixed at 140 groups. The size of the inertial sensing data input is fixed at 130 groups. Therefore, if the size of array A is smaller than 140, it is necessary to obtain the inertial sensing data of size 140 by linear interpolation. Therefore, if the size of array A is smaller than 140, an array of size 140 needs to be obtained as the input by linear interpolation. Linear interpolation is a one-dimensional data interpolation method; it is through a one-dimensional data sequence of the data points interpolated with the left and right two neighboring points. The estimation of the value is performed by assigning the weight of these two points according to their distance to the point to be inserted.

3.2 Finger trace tracking and gesture classification module

We divide the letter recognition into two sub-processes: fingertip letter trajectory tracking and trajectory classification. Firstly, based on the user's handwritten letters, and then classify the user's handwritten letters by combining the original inertial sensing data to track the trajectory of the user's fingertip during handwriting. When tracking the fingertip trajectory of the user's handwritten letters, we need to calculate the position of the key joint point first, and then combine the position of the key joint point with the sliding filtering algorithm to the trajectory of the index fingertip is deduced based on the position of the key nodes and the sliding filter algorithm [9, 10].

It is first necessary to determine which joint points on the palm are the critical ones. After observation, we found that users mainly use their index finger to write letters, so we selected the tip of the index finger, index finger metacarpal, and wrist bone as the key joints. After identifying the key joints, to track the trajectories of these joints, the temporal position coordinates of these joints need to be obtained first. We propose a Transformers and CNN-based method, named TranC, to calculate the timing position coordinates of key joints. The TranC neural network takes pre-processed inertial sensing data as input. The architecture of the TranC is illustrated in figure 1. The size of the input layer is 140*12. For the first three convolutional layers, we make use of 16 filters with the size of 3*3. For the fourth convolutional layer, we make use of 8 filters with the size of 3*3. However, there is a problem in that the above convolutional layers focus on extracting local features while ignoring the importance of the global gesture information. Therefore, to address the problem, we leverage an additional Transformer to further capture the global gesture information.

3.3 Model Training

The data needed in the training process are the inertial sensing data generated by the wrist when the user writes letters by hand, which can be collected by the IMU of the smartwatch, and the ground truth of the fingertip trajectory data, the spatial position coordinates of the 3 key joint points in 140 sets of time series. In this paper, the inertial sensing data collected by the IMU is defined as I. The ground truth of the fingertip trajectory captured by Leap Motion is defined as G. Then, the training objective function can be defined as:

$$\text{loss} = \min_{\text{TranC}} \sum\nolimits_{(\text{U,V})} \text{L}\left(\text{TranC}(\text{I}), \text{G}\right) \tag{4}$$

The Eq. 4 can be further rewritten as:

$$\text{loss} = \text{C}\left(1^{1}_{\text{mcp}}, 1^{\text{G}}_{\text{mcp}}\right) \cdot \text{C}\left(1^{\text{I}}_{\text{if}}, 1^{\text{G}}_{\text{if}}\right) \cdot \sum_{\text{r=1}}^{140} \sum_{\text{s=1}}^{3} \text{D}^2\left(\text{TranC}_{\text{rs}}, \text{G}_{\text{rs}}\right) \tag{5}$$

where l_{if} denotes the length of the index finger calculated from the position of the index finger tip and the index finger metacarpal degree. l_{mcp} is the length of the metacarpal bone calculated from the position of the metacarpal bone of the index finger and the wrist carpal bone. 1^{G}_{if} and $1^{\text{G}}_{\text{mcp}}$ are the ground-truth labels.

Table 1: Experimental results

Models	Gesture Dataset	
	Acc	F1
RF-finger	92.1%	91.1%
UWB-based	90.6%	89.7%
Widraw	90.4%	88.4%
SoundWave	91.0%	90.2%
TranC	93.1%	91.3%

4 EXPERIMENTS

In this section, we conduct experiments to validate the performance of the proposed model.

4.1 Experimental Settings

We conduct experiments in the environment of the PyTorch framework. Additionally, all the neural network-based models use Adam as the optimiser with a learning rate of 1e-5. Besides, the epoch for the neural network-based models is set to 100 with an early stopping mechanism. The number of hidden states is set to 128. The number of heads in the multi-head attention part is set to 8. We run the models for 3 times, and took the average of the results as the final results. In this experiment, the performance of the gesture model recognition system is evaluated using two evaluation metrics: Accuracy and Macro-F1. If an instance is a positive class and is predicted to be positive, it is a true class TP (True Positive); if an instance is a positive class but is predicted to be negative, it is a false negative class FN (False Negative); if an instance is a negative class but is predicted to be positive, it is a false positive class FP (False Positive); if an instance is a negative class and is predicted to be negative, it is a true negative class TN (True Negative).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

4.2 Datasets

We conduct experiments on a benchmark dataset for the task of human gesture recognition. The dataset contains 15,000 samples with ground-truth labels.

4.3 Experimental Results

In order to verify the superiority of the proposed gesture recognition methods in gesture classification, this paper compares the accuracy and the macro-F1 of the proposed TranC with several other types of gesture recognition methods introduced in related works. The experimental results are illustrated in the table 1. From the table, we can observe that the proposed method is superior to other state-of-the-art baselines by a significant margin.

5 CONCLUSION

Most of these methods are based on the convolutional neural network, which extracts deep features from a local manna. However, this way impedes the models from learning global information. To solve the problem, in this paper, we propose a novel method named TranC that leverages the attention mechanism to capture global information about human gestures. We conduct experiments on a benchmark, and the experimental results demonstrate that our proposed method is superior to the other baselines.

REFERENCES

[1] Patsadu, Orasa, Chakarida Nukoolkit, and Bunthit Watanapa. "Human gesture recognition using Kinect camera." 2012 ninth international conference on computer science and software engineering (JCSSE). IEEE, 2012.
[2] Liao, Shangchun, *et al.* "Multi-object intergroup gesture recognition combined with fusion feature and KNN algorithm." Journal of Intelligent & Fuzzy Systems 38.3, 2020: 2725-2735.
[3] Oh, Juhee, Taehyub Kim, and Hyunki Hong. "Using binary decision tree and multiclass SVM for human gesture recognition." 2013 International Conference on Information Science and Applications (ICISA). IEEE, 2013.
[4] Carfi, Alessandro, *et al.* "Online human gesture recognition using recurrent neural networks and wearable sensors." 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, 2018.
[5] Xu, Dan, *et al.* "Online dynamic gesture recognition for human robot interaction." Journal of Intelligent & Robotic Systems 77.3, 2015: 583-596.
[6] Fang, Bin, *et al.* "3D human gesture capturing and recognition by the IMMU-based data glove." Neurocomputing 277, 2018: 198-207.
[7] Utsumi, Akira, and Jun Ohya. "Multiple-hand-gesture tracking using multiple cameras." Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). Vol. 1. IEEE, 1999.
[8] Khan, Salman, *et al.* "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s, 2022: 1-41.
[9] Ahmed, Hasmath Farhana Thariq, Hafisoh Ahmad, and C. V. Aravind. "Device free human gesture recognition using Wi-Fi CSI: A survey." Engineering Applications of Artificial Intelligence 87, 2020: 103281.
[10] Lin, Hsien-I., Ming-Hsiang Hsu, and Wei-Kai Chen. "Human hand gesture recognition using a convolution neural network." 2014 IEEE International Conference on Automation Science and Engineering (CASE). IEEE, 2014.