# Human action recognition from inertial sensors with Transformer

Trung-Hieu Le[1,3], Thanh-Hai Tran[1], Cuong Pham[2]

[1] School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
E-mail: hieult@dainam.edu.vn
[2] Department of Computer Science, Posts and Telecom Institute of Technology, Hanoi, Vietnam
[3] Dainam University, Hanoi, Vietnam

2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR) | 978-1-6654-7410-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/MAPR56351.2022.9924794

***Abstract*—Human action recognition is an attractive research topic because it opens many practical applications such as healthcare, entertainment or robot interaction. Hand gestures in particular are becoming one of the most convenient means of communication between humans and machines. In this study, transformer model - a deep learning neural network developed primarily for the natural language processing and vision tasks, is investigated for analysis of time-series signals. The self-attention mechanism inherent in the transformer expresses individual dependencies between signal values within time series. As a result, it can boost the performance of state-of-the-art convolutional neural networks in terms of memory requirement and computational times. We evaluate the proposed method on three published sensor datasets (CMDFALL, C-MHAD and DaLiAc) and showed that the proposed method achieves better performance than conventional ones, specifically on the $S_3$ group in the CMDFall data set, the F1 Score is 19.04 % higher than that of the conventional method. On C-MHAD dataset, the accuracy is up to 99.56 %. The results confirms the role of transformer models for human activity recognition.**

***Index Terms*—Human activity recognition, Classification, inertial sensor, Convolutional Neural Networks, Transformers, Sequence Analysis.**

## I. Introduction

Recognizing human actions and human hand gestures in particular has significantly attracted the research community as it opens a wide range of practical applications such as healthcare, entertainment or robot interaction [1]. To recognize a human action, that action must be first captured by some sensors (e.g. cameras, accelerometers, gyroscopes) and then fed into a recognition model to predict the action label. While cameras have been widely used in literature because it can provide rich information about scene [2], inertial sensors (i.e. accelerometer and gyroscope) has its own advantages such as the low cost and compact size to be integrated into smart wearable devices [3], [4]. In addition, such inertial sensors can avoid user privacy and scale to any environment. As a result, many works have utilized inertial sensors for human action recognition [3].

Motion-based action recognition methods can be categorized into two groups. The first group tried to extract handcrafted features (e.g. statistical features) from signal then passed them though a classifier such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), decision tree (DT), etc [4]. The second group bases on learning hidden features from a huge dataset of annotated actions. In recent years, CNN models have shown their impressive performance on many tasks, including human action recognition from multi-modal data.

To process time series signal, conventional recurrent neural networks such as Long-Short Term Memory (LSTM), GRU can be employed. Variations of LSTM such as bi-directional LSTM, LSTM with Batch normalization showed better performance than the original ones [5], [6]. Recently, transformer models have been introduced and show its out performance in the task of natural language processing [7]. Some transformer models have also applied for vision-based tasks [8], [9]. However, very little works on human action recognition using inertial sensors employed transformer models.

In this paper, we make use transformer models to investigate its performance on action recognition using inertial sensors. The transformer model relied upon attention mechanism to find time series correlations between features and allows for large parallelism of time series computations, which is different from lattice cyclic neurons repeat in time series. Another advantage of transformers is longer path lengths between objects in time series, allowing more accurate learning of context in long time series, a claim by Vaswani et al [7]. Besides using Transformers, we investigate how the size and stride of sliding window impacts the performance of transformer models. We evaluate the studied model on 03 datasets that have been published to the research community. The results showed that transformer model achieves better results than the method initial on the three evaluated datasets.

In summary, the contribution of this study is two fold. First, we study the transformer model for the problem of human action recognition from inertial sensors (i.e acclerometers and gyroscopes). We consider the role of separated modality in transformer model then their combination. Second, we evaluate the effectiveness of the model on three public datasets (execution time, recognition accuracy) according to the size and stride of sliding window. The rest of the paper is structured as follows. We summarize relevant works in section II, then we introduce our framework in section III. Experiments and conclusions are presented in sections VI and V respectively.

## II. Related works

### A. Sensor mounting positions

Common wearable sensors for human action recognition are accelerometers or gyroscopes due to their low-cost and compact size, easy to wear without disturbing the users. Depending on the activities of interest, the mounting position of these sensors may be different. Akin et al [10] presented an overview of studies using wearable accelerometers. Common mounting locations of accelerometers are thighs, hips, chest, wrists, arms, ankles, shoe soles or at all major joints of the body. The activities measured by the accelerometers / gyroscope can take place indoors or outdoors, from cooking, personal activities (brushing teeth), walking and climbing stairs to running, playing ball, sports, mountain climbing. In a study assisting the elderly by Stylianos Paraschiakos [11], accelerometers were mounted at various locations. From a modelling perspective, it has been proven that HAR model's performance is highly dependent on the sensors' body position in Bao et al. [12].

### B. Human action recognition from inertial sensors

Some studies show that human action recognition from wearable sensors can be highly effective in man-machine interaction applications such as robotic arm control [13], indoor device controlling [14] and robot controlling [15] with hand gestures. In recent times, deep neural networks have been strongly developed and obtained good results in many problems, deep learning techniques have also been widely applied to the problem of recognizing human activities based on the inertial sensor.

In the study of Alsheikh and his colleagues [16] , the authors converted the acceleration signal to the frequency domain (spectrogram) before being fed into the Deep Belief Network (DBN). The results show that using a deep neural network gives better results than using traditional methods. In 2018, Shakya [17] conducted a comparative study on traditional architectures and deep learning networks (CNNs) and RNNs (recurrent neural networks). Experiments showed that CNN gives better results than RNN and traditional methods. Tran et al [2] proposed a method to detect falls over time from the accelerometer. Khaertdinov et al. proposed a Network that combines three sets of LSTMs and focus blocks for sensor-based action recognition [18]. Nguyen et al propose 1D-CNN biLSTM architecture for recognizing 18 gestures using motion sensors [19]. Li et al [20] joined 2D convolution layers followed by a BiLSTM layer. The BiLSTM allows the time series to move in two directions, from the past into the future and from the future into the past.

In our study, we deal with the application of deep neural networks directly to the normalized time series of signals from sensors. This study uses an alternative approach to time series processing based entirely on the attention mechanism, known as the transformer. The transformer model directly focuses on using the attention mechanism to find time series correlations between features and allows for large parallelism of time series computations, which is different from lattice cyclic neurons in time series. Calculation speed, as well as prediction accuracy, are important factors in working with human activities, where predictions can be made directly on a wearable device. Another advantage of the transformer is the longer path lengths between objects in the time series, allowing for more accurate learning of the context in the long time series [7]. The sequential method is used in the prediction of operations, where all the time steps from the transformer output are considered and the specified operation is assigned to them. In this way, it is possible to assign an operation to each time step that the user has taken when measuring values directly from the sensor-mounted wearable.

## III. Method for human action recognition using transformer

### A. Proposed framework

Previous studies often refer to important neural network architectures such as convolutional neural networks (CNNs) and feedback neural networks (RNNs). While CNNs can easily be parallelized at a single layer, they are not capable of capturing variable-length sequence dependencies. Besides, RNNs are capable of capturing routing information far apart in a variable-length sequence, but cannot be parallelized within a sequence. To combine the advantages of CNN and RNN, Vaswani et al [7] designed a new architecture using centralization. This architecture, called Transformer, parallelizes by learning the feedback sequence with a centralized mechanism, and also encodes the position of each element in the sequence. The results show a compatible model with a significantly shorter training time. Before the Transformer model was introduced, time series processing problems were often based on supervised, cross-attention mechanisms as studied by Jang et al. [21]. However, for the Transformer model it is shown that attention mechanisms do not need to use optimal parameters to achieve performance, and that the Transformer model can more easily combine parallel threads with an accumulative neural network.

In this paper, we employ transformer model for human action recognition from inertial sensors. The proposed recognition framework is depicted in Fig. 1. Given a sample of data $\mathbf{S} \in \mathbb{R}^{k \times 6}$ (3 values of accelerometer and three values of gyroscope), k is the number of data samples (window size). The problem of human action recognition is considered as sequence-to-one where input is a sequence of sensor measurements and output is the label of action. The transformer will be used to extract latent embedding from raw data and a Multilayer Perceptron (MLP) can be employed for classification. Fig. 1 shows a framework of three steps: pre-procesing, transformer encoding and classification. In the following, we will describe in detail each of the steps.

### B. Pre-processing

*1) Data segmentation::* The raw data will be loaded into the framework through the data loader. The data loader will be responsible for removing redundant or missing data gestures

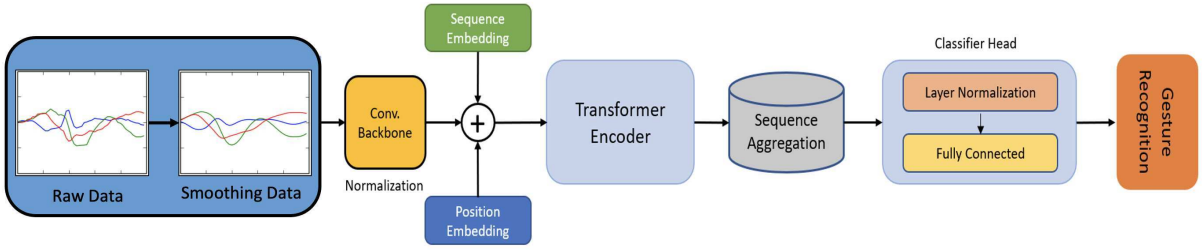


Fig. 1. The transformer model for action recognition.

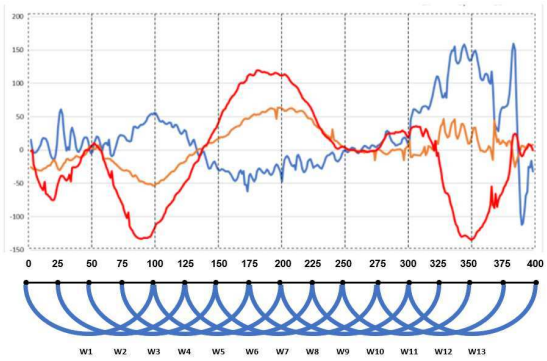


Fig. 2. Sliding window method

(noise filtering) and then assigning data labels. The data loader outputs data from the accelerometer and gyroscope. Let $(a_x, a_y, a_z)$ and $(g_x, g_y, g_z)$ denote the data from accelerometer and gyroscope sensors respectively.

We use sliding window technique with 25 % overlap for all experimented dataset as shown in Fig.2. The signal in each window is considered to contain a specific gesture. In DaLiAc dataset [22], we selected the windows size as 5s, as used in [23]. In CHMAD [24] and CMDFALL [2] datasets, we experiment with different window sizes to choose the best one.

*2) Noise removing:* The data needs to be smoothed before put into model Transformer to maximize the performance of the method. In this paper, we utilize the following smoothing function:

$$x_i = \frac{(x_{i-2} + x_{i-1} + x_i + x_{i+1} + x_{i+2})}{5} \tag{1}$$

where $x_i$ is the data signal at the time $i$. This formula is applied for all channels of accelerometer and gyroscope data and also the magnitudes of these two signals.

### C. Embedding raw data

To encode the raw sequence, we utilize a convolutional neural network which consists of a series of four 1D convolutions with GELU (Gaussian Error Linear Units) activation. This step embeds the raw data in a higher dimension $d$ generating a latent embedding $\mathbf{E_S} \in \mathbb{R}^{k \times d}$ [25]. A class token $\mathbf{C} \in \mathbb{R}^d$ is prepended to the embedded sequence $\mathbf{E_S}$. In addition, for each position $P_i$ in the sequence, an embedding $\mathbf{E_{P_i}} \in \mathbb{R}^d$ is learned and added to the latent sequence representation. As a consequence, the initial input $Z_0$ to the Transformer Encoder is as follows:

$$\mathbf{Z_0} = [\mathbf{C}, \mathbf{E_s}] + \mathbf{E_P} \in \mathbb{R}^{t \times d} \tag{2}$$

where t = k + 1.

### D. Transformer encoder

In the Transformer model, the attention mechanism is an important part. In which, for each attention head belonging to different attention heads, it can show that there are many definitions of correlation and relevance [26]. The rules that map queries and sets of key-value pairs to an output are Multi-head attention [7]. The Total Output of the network is the weighted sum of the values V, Q, and K. Where the Weight is assigned to each value (V) based on the calculation of the compatibility function from the query (Q) and (K) is the corresponding key. The dot product of the query and all keys will be calculated, then the softmax function is included to normalize the obtained parameters, and then multiplied by the values. Softmax is the basis for prediction given classifiers, where its output values determine the probabilities of distinct items.

The encoder architecture is adopted from [7]. It composes of $L$ layers, each layers consists of a self multi-head attention (MHA) layer and an Multi-layer Perceptro (MLP). In the implementation, the MLP contains two fully connected (FC) layers with a hidden dimension of 2 $d$ and GELU non-linearity. Multi-head attention operation is the core of the Transformer model. We investigate three sequences of length $t$ and dimension $d$: a query $\mathbf{Q} \in \mathbb{R}^{t \times d}$, a key $\mathbf{K} \in \mathbb{R}^{t \times d}$ and a value $\mathbf{V} \in \mathbb{R}^{t \times d}$, each head $h$ computes a weighted aggregation of V with $Q$:

$$\mathbf{h}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q_h}\left(\mathbf{K_h^T}\right)}{\sqrt{d}}\right)\mathbf{V_h} \in \mathbb{R}^{t \times d'} \tag{3}$$

where:

$$\mathbf{Q_h} = \mathbf{Q}\mathbf{W_h^Q} \in \mathbb{R}^{t \times d'} \tag{4}$$

$$\mathbf{K_h} = \mathbf{K}\mathbf{W_h^K} \in \mathbb{R}^{t \times d'} \tag{5}$$

$$\mathbf{V_h} = \mathbf{V}\mathbf{W_h^V} \in \mathbb{R}^{t \times d'} \tag{6}$$

where $d' = \frac{d}{n_h}$ and $n_h$ is the number of heads. The matrices $\mathbf{W_h^Q}, \mathbf{W_h^K}, \mathbf{W_h^V} \in \mathbb{R}^{d \times d'}$ are linear projections from $d$ to $d'$.

The outputs of query value Q, key K, value V and softmax function are connected according to the size of the channel. The resulting update representation is a weighted sum of the sequence at each position, based on the relative importance of all other positions. Multi-head attention softmax itself, Q, K and V are considered to be the same sequence.

Each layer $l \in [1, L]$ for the Transformer Encoder, the calculations are done by passing the input through Layer Normalization (LN) [27] before each module and adding it with the remaining connections:

$$\mathbf{Z}'_1 = sMHA\left(LN\left(\mathbf{Z}_{1-1}\right)\right) + \mathbf{Z}_{1-\mathbf{1}} \in \mathbb{R}^{t \times d} \tag{7}$$

$$\mathbf{Z}_1 = MLP\left(LN\left(\mathbf{Z}'_1\right)\right) + \mathbf{Z}'_1 \in \mathbb{R}^{t \times d} \tag{8}$$

The output of the Transformer Encoder in place of the class token represents the transient synthesis of the input sequence with the formula:

$$\mathbf{Y_C} = \mathbf{Z_L}[\mathbf{0}] \in \mathbb{R}^d \tag{9}$$

$\mathbf{Y_C}$ is provided as input to the classifier head, consisting of LN and Fully connected layers with non-linear GELU and Dropout, reducing the dimension to $\frac{d}{4}$. The second Fully connected layer maps $\frac{d}{4}$ with the class number. A log SoftMax is applied on the output vector along with negative log loss to learn the multi-label classification task [25].

## IV. Experiment

### A. Datasets

We use three available datasets to evaluate the studied method, including CMDFALL [2], C-MHAD [24] and DaLiAc [22]. The data of each gesture/activity used will include six signals: three signals in 3 $x$, $y$, $z$ axes of the accelerometer sensor and three signals in $x$, $y$, $z$ axes of the gyroscope sensor. The details of each data set are described as follows:

- **CMDFALL [2]:** This is a rather large dataset collected from 50 people wearing two sensors at wrist and waist position. The data set includes nine normal activities (such as walking, lying on the bed, sitting down in a chair, etc.) and 11 abnormal movements (such as falling on the back, falling side, staggering, slipping ...). The sampling frequency of the data set is 50 Hz.
- **C-MHAD [24]**: The continuous dataset contains five hand gestures for controlling smart TVs performed by 12 subjects (10 males and 2 females). Ten continuous streams of video and inertial data, each lasting for 2 minutes, were captured for each subject. The inertial signals in this dataset consist of 3-axis acceleration signals and 3-axis angular velocity signals which were captured by the commercially available Shimmer3 wearable inertial sensor at a frequency of 50Hz on a laptop via a Bluetooth link.
- **DaLiAc [22]**: In the DaLiAc dataset, 4 inertial measurement units each consisting of a triaxial accelerometer and a triaxial gyroscope were used. The sensors were placed on the right hip, chest, right wrist, and left ankle. The sampling rate was 204.8 Hz. The dataset includes 19 subjects (8 female and 11 male). Each subject had to perform 13 daily life activities.

TABLE I
INFORMATION ABOUT DATASETS

| CHARACTERISTICS | CMDFALL | CHMAD | DaLiAc |
|---|---|---|---|
| Types of activities | Daily activities | Hand gesture | Daily activities |
| Number of gestures | S2: 6 S3: 20 | 5 | 13 |
| Data types | Acc & Acc | Acc & Gyro | Acc & Gyro |
| Sampling rate (Hz) | 50 | 50 | 204, 8 |
| Total activities /gestures obtained | 2353 | 1018 | 266 |
| Average number of samples/activity) | 305, 7 | 102, 92 | 17619.64 |
| Split training, test | Even ID train Odd ID test data | K − Fold Cross (K = 10) | Train test split (30% test) |

### B. Experimental results

In the experiment, we will compare the use of different window sizes (w = 64, 96, 128) for each dataset. Then we will analyze the performance achieved with the selected window size applied to the Transformer model. In addition, we also compare with the results of previous publications to confirm the performance of the transformer model.

- **CMDFALL dataset:** With this dataset, we evaluate our method for two groups of activities $S_2$ (6 activities) and $S_3$ (20 activities) [2]. Tab. II shows the results obtained by using window size w = 64, 96, 128. We found that the Transformer model achieved a very high accuracy of 64.99 % with w = 128 on the $S_2$ group. The same conclusion is made for the group $S_3$. This group has many classes that cause classification confusion. With the right window size, the accuracy increases from 58.07 % (w = 64) to 60.19 % (w = 128). This shows that suitable selection of the window size may significantly improve the recognition accuracy.
  We also compare the transformer model with the method introduced in [2]. Fig. 3 shows F1 score on two groups $S_2$ and $S_3$ of CMDFALL dataset produced by [2] method and our method using a Transformer model with the window size of 128 (w = 128). The results show that the F1 score of our method increased compared to the results of the previous study. On the group $S_2$, the F1 score increased only slightly, while it increased significantly on the group $S_3$ by 19.04 %.
- **C-MHAD dataset**: Similar to the CMDFALL dataset, the accuracy of our proposed method on the C-MHAD dataset is significantly improved. At w = 64, the accuracy achieves 88.86 %. Especially with w=96 and w=128 the accuracy increases to 97.93% and 99.56% (i.e. from 8 % to more than 11 %). It shows that the recognition

TABLE II
COMPARISON OF RECOGNITION ACCURACY (%) ON THE GROUPS $S_2$ AND THE GROUP $S_3$ OF CMDFALL DATASET

| Group | Window size | Accuracy |
|---|---|---|
| S2 | 64 | 64.33 |
| | 96 | 64.71 |
| | 128 | 64.99 |
| S3 | 64 | 58.07 |
| | 96 | 58.17 |
| | 128 | 60.19 |

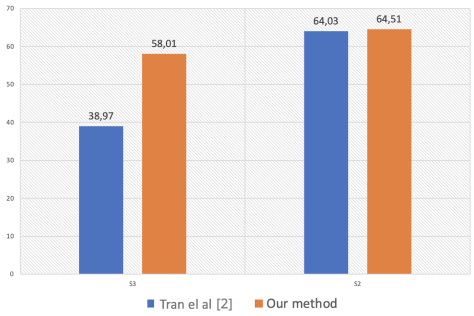


Fig. 3. F1 score (%) on two groups of CMDFALL dataset

is improved with additional use of the corresponding window size.

- **DaLiAc dataset**: Tab.IV shows the results obtained by transformer model with different values of window size (w = 64, 96, 128). We observe that on this dataset, our method only slightly improves the recognition rate. The reason is that this dataset contains daily life activities which usually do not have a clear starting and ending time. In addition, there is no significant difference between segments inside the activity. The detailed results are as follows: w = 128 achieves the best accuracy of 96.86 %, the following are w = 96 with an accuracy of 85.56% and w = 64 with an accuracy of 82.03%.
We compare the performance of some existing works on DaLiAc dataset [22], [28] and [29]. Comparison results show that our results are slightly higher than method of Chen et al [28] and 7 to 9% more than the other 2 methods. Fig. 7 shows the comparison result.

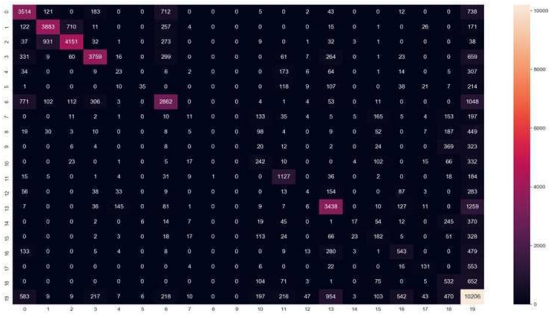

Fig. 4. Confusion matrix on $S_3$ of CMDFALL dataset using Transformer with window size w = 128.

TABLE III
COMPARISON OF RECOGNITION ACCURACY (%) ON C-MHAD DATASET

| Dataset | Window size | Accuracy |
|---|---|---|
| C-MHAD | 64 | 88.86 |
| | 96 | 97.93 |
| | 128 | 99.56 |

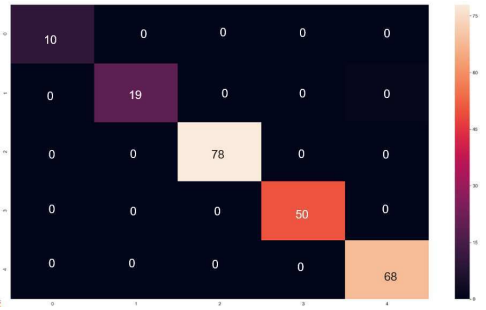


Fig. 5. Confusion matrix on CHMAD Dataset using Transformer with window size w = 128

TABLE IV
RECOGNITION ACCURACY (%) ON DALIAC DATASET

| Dataset | Window size | Accuracy |
|---|---|---|
| DaLiAc | 64 | 82.03 |
| | 96 | 85.56 |
| | 128 | 96.86 |

## V. CONCLUSION

This paper presents a deep learning framework for gesture recognition. The proposed approach employed Transformer. This study is one of the most recent to study the Transformer architecture used for gesture recognition based on accelerometer and gyroscope data. Three types of accelerometer datasets, with activities/gestures collected by the users, were used for objective evaluation: 1) CMDFALL, generated from 20 activities in the research environment, 2) CHMAD, with 5

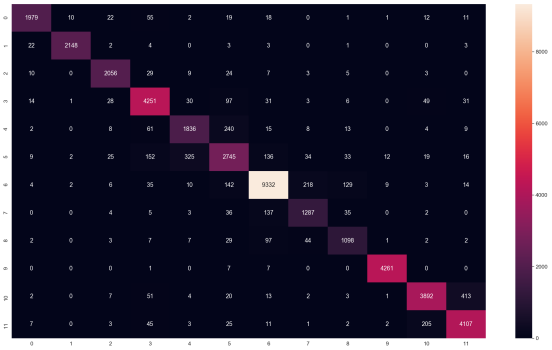

Fig. 6. Confusion matrix on DaLiAc Dataset using Transformer with window size w = 128.

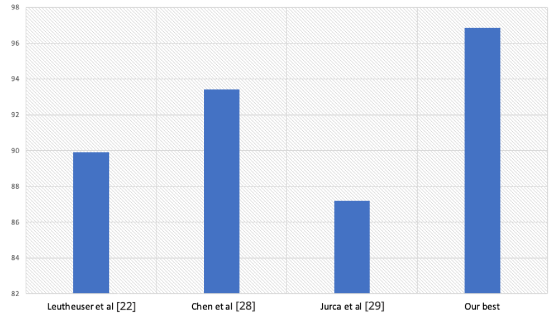


Fig. 7. Recognition accuracy (%) on DaLiAc dataset

gestures to control smart TV, and, 3) DaLiAc to recognize the human activity. The results showed that the Transformer architecture gives better results than the original methods and is a prerequisite for carrying out further studies. In the future, the Transformer model will be tested by us on an extended dataset, ideally using various sensor data that we collect and publish ourselves. After achieving the desired results, it will be applied first to build an application for controlling smart homes by hand gestures. In addition, they will serve as a springboard for further useful applications involving direct human assistance.

## ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-20-1-4053.

## REFERENCES

[1] Feiyu Chen, Jia Deng, Zhibo Pang, Majid Baghaei Nejad, Huayong Yang, and Geng Yang. Finger angle-based hand gesture recognition for smart infrastructure using wearable wrist-worn camera. *Applied Sciences*, 8(3):369, 2018.

[2] Thanh-Hai Tran, Thi-Lan Le, Dinh-Tan Pham, Van-Nam Hoang, Van-Minh Khong, Quoc-Toan Tran, Thai-Son Nguyen, and Cuong Pham. A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1947–1952. IEEE, 2018.

[3] Cuong Pham, Nguyen Ngoc Diep, and Tu Minh Phuong. A wearable sensor based approach to real-time fall detection and fine-grained activity recognition. *Journal of Mobile Multimedia*, pages 015–026, 2013.

[4] Trung-Hieu Le, Thanh-Hai Tran, and Cuong Pham. The internet-of-things based hand gestures using wearable sensors for human machine interaction. In *2019 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE, 2019.

[5] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[6] Tahmina Zebin, Matthew Sperrin, Niels Peek, and Alexander J Casson. Human activity recognition from inertial sensor time-series using batch normalized deep lstm recurrent networks. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 1–4. IEEE, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[8] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[9] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[10] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *23th International conference on architecture of computing systems 2010*, pages 1–10. VDE, 2010.

[11] Stylianos Paraschiakos, Ricardo Cachucho, Matthijs Moed, Diana van Heemst, Simon Mooijaart, Eline P Slagboom, Arno Knobbe, and Marian Beekman. Activity recognition using wearable sensors for tracking the elderly. *User Modeling and User-Adapted Interaction*, 30(3):567–605, 2020.

[12] Ling Bao and Stephen S Intille. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*, pages 1–17. Springer, 2004.

[13] Shriya A Hande and Nitin R Chopde. Implementation of gesture control robotic arm for automation of industrial application. 2020.

[14] Yen-Cheng Chu, Yun-Jie Jhang, Tsung-Ming Tai, and Wen-Jyi Hwang. Recognition of hand gesture sequences by accelerometers and gyroscopes. *Applied Sciences*, 10(18):6507, 2020.

[15] Saleem Ullah, Zain Mumtaz, Shuo Liu, Mohammad Abubaqr, Athar Mahboob, and Hamza Ahmad Madni. Single-equipment with multiple-application for an automated robot-car control system. *Sensors*, 19(3):662, 2019.

[16] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. Deep activity recognition models with triaxial accelerometers. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[17] Sarbagya Ratna Shakya, Chaoyang Zhang, and Zhaoxian Zhou. Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data. *Int. J. Mach. Learn. Comput*, 8(6):577–582, 2018.

[18] Bulat Khaertdinov, Esam Ghaleb, and Stylianos Asteriadis. Deep triplet networks with attention for sensor-based human activity recognition. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2021.

[19] Khanh Nguyen-Trong, Hoai Nam Vu, Ngon Nguyen Trung, and Cuong Pham. Gesture recognition using wearable sensors with bi-long short-term memory convolutional neural networks. *IEEE Sensors Journal*, 21(13):15065–15079, 2021.

[20] Yong Li and Luping Wang. Human activity recognition based on residual network and bilstm. *Sensors*, 22(2):635, 2022.

[21] Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, Sang-ug Kang, and Jong Wook Kim. Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. *Applied Sciences*, 10(17):5841, 2020.

[22] Heike Leutheuser, Dominik Schuldhaus, and Bjoern M Eskofier. Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art algorithms using a benchmark dataset. *PloS one*, 8(10):e75196, 2013.

[23] Jianchao Lu, Xi Zheng, Michael Sheng, Jiong Jin, and Shui Yu. Efficient human activity recognition using a single wearable sensor. *IEEE Internet of Things Journal*, 7(11):11137–11146, 2020.

[24] Haoran Wei, Pranav Chopada, and Nasser Kehtarnavaz. C-mhad: Continuous multimodal human action dataset of simultaneous video and inertial sensing. *Sensors*, 20(10):2905, 2020.

[25] Yoli Shavit and Itzik Klein. Boosting inertial-based human activity recognition with transformers. *IEEE Access*, 9:53540–53547, 2021.

[26] Iveta Dirgová Luptáková, Martin Kubovčík, and Jiří Pospíchal. Wearable sensor-based human activity recognition with transformer model. *Sensors*, 22(5):1911, 2022.

[27] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[28] Ye Chen, Ming Guo, and Zhelong Wang. An improved algorithm for human activity recognition using wearable sensors. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, pages 248–252. IEEE, 2016.

[29] Roxana Jurca, Tudor Cioara, Ionut Anghel, Marcel Antal, Claudia Pop, and Dorin Moldovan. Activities of daily living classification using recurrent neural networks. In *2018 17th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, pages 1–4. IEEE, 2018.