SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Action Transformer: Model Improvement and Effective Investigation with MPOSE2021 and MSR Action 3D Datasets

Khac-Anh Phu[1], Van-Dung Hoang[2,*], Van-Tuong-Lan Le[3]

[1] Faculty of Information Technology, University of Sciences, Hue University, Hue City, 530000, Vietnam
[2] Faculty of Information Technology, HCMC University of Technology and Education, Ho Chi Minh City, 700000, Vietnam
[3] University of Sciences, Hue University, Hue city, 530000, Vietnam

ABSTRACT

The AcT (Action Transformer) model has shown promising results in action recognition tasks. However, achieving high accuracy in complex and dynamic action sequences remains a challenge. In this paper, we present an approach to improve the accuracy of the AcT model by increasing the model's training complexity, validated on the MPOSE2021 and MSR Action datasets. Our method enhances the AcT model by incorporating a multi-level feature fusion technique. We introduce additional convolutional and pooling layers to capture more detailed spatial and temporal information from the input data. This increases the model's ability to discriminate between subtle action variations and improves its accuracy in recognizing complex actions. We evaluate the effectiveness of our proposed approach through extensive experiments on the MPOSE2021 and MSR Action datasets. The results demonstrate that our enhanced AcT model achieves significantly improved accuracy compared to the baseline AcT model and outperforms existing state-of-the-art methods. Our method effectively captures the intricacies of complex actions and provides more accurate predictions.

Keywords:
Action transformer; human action recognition; skeleton data; deep learning

1. Introduction

Human Action Recognition (HAR) is a significant research area in the fields of computer vision and artificial intelligence. The task of HAR is to identify and classify human actions from input data such as images or videos. In recent years, numerous HAR methods have been proposed and developed. However, achieving high accuracy in action recognition remains a challenge [1]. An important factor in improving HAR accuracy is the input data [2]. Among the commonly used input data types in HAR, skeleton data has been proven to be highly effective and was first introduced in [3]. Skeleton data is an abstract representation of humans, where each skeleton joint represents a specific position on the body and the relationships between them. Utilizing skeleton data brings numerous significant benefits.

* *Corresponding author.*
E-mail address: dunghv@hcmute.edu.vn

https://doi.org/10.37934/araset.62.1.7689

Firstly, skeleton data can reduce the dependence on external factors such as lighting, background, and surrounding environment. This is because skeleton data focuses solely on the positions and basic relationships of the human body, disregarding irrelevant factors. Therefore, utilizing skeleton data can improve the accuracy of HAR in conditions with low lighting, complex backgrounds, or unfavorable environments. Secondly, skeleton data helps reduce computational costs and storage requirements. Compared to full-image or video data, skeleton data is more compact and has lower dimensionality. This makes the processing and analysis of data more efficient, while also reducing the demands on computational resources and storage.

The AcT (Action Transformer) model is a powerful approach for action recognition, built upon the foundation of the Vision Transformer architecture [4], and introduced by the authors in their study [5] on short-term action recognition using skeleton data. The Transformer architecture [6], which the AcT model is based on, is one of the most significant advancements in the field of deep learning for natural language processing (NLP) in recent years. The self-attention mechanism in the Transformer architecture, with multiple heads, has been proven effective for various tasks beyond NLP, such as image classification [4, 7, 8], image super-resolution [9, 10], and speech recognition [11]. However, to enhance the performance and generalization capabilities of the model, we introduce the AcTv2 model, which is an improved version of the AcT model by incorporating additional important layers into the original AcT model.

In our study, we conducted experiments on the AcTv2 model using two important datasets, namely MPOSE2021 and MSR Action3D. By utilizing the MPOSE2021 and MSR Action 3D datasets, we had the opportunity to evaluate the performance of the AcTv2 model in recognizing various actions. Through the experimental process, we collected data and evaluated the accuracy of the model on these datasets.

The experimental results have demonstrated that the AcTv2 model, with its improvements, achieved significantly higher accuracy compared to the original model. This increase in accuracy highlights the effectiveness of the enhancements we applied to the AcT model. It indicates that the potential and capabilities of the AcT model have been enhanced through the use of new methods and datasets.

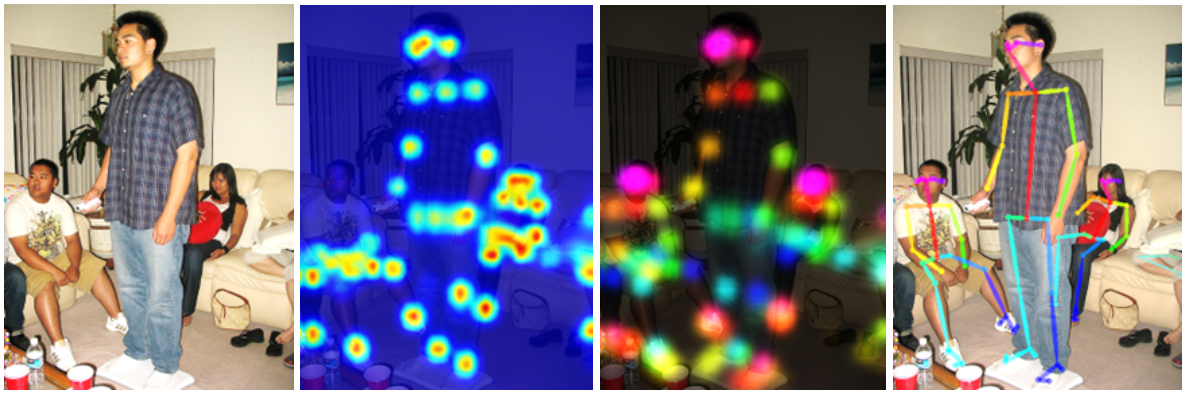
In summary, our research contributes to two important aspects in the field of action recognition:

i. Firstly, we proposed the AcTv2 model, an improved version of the AcT model, aiming to enhance the accuracy in recognizing actions on the MPOSE2021 dataset. Through experiments on the MPOSE2021 dataset, we demonstrated that the AcTv2 model achieved higher accuracy compared to the original AcT model. This improvement was evaluated by comparing the accuracy results of the two models on the MPOSE2021 dataset. The experimental results showed that the AcTv2 model helps recognize actions with better accuracy and enhances the ability to recognize complex action sequences in this dataset.
ii. Secondly, we conducted experiments with the AcT model on the MSR Action 3D dataset to evaluate its effectiveness compared to other algorithms in the field of action recognition. The experimental results showed that the AcT model outperformed and achieved higher accuracy than the majority of algorithms previously published on the MSR Action 3D dataset. Based on these results, we applied the improved version, AcTv2, with higher accuracy, to perform action recognition on the MSR Action 3D dataset. The obtained results demonstrated a significant improvement in accuracy with AcTv2 compared to the AcT model.

Overall, our research makes significant contributions to the development and improvement of action recognition methods and holds potential for broad applications in fields related to artificial intelligence and computer vision.

2. Related Work

As presented in Section I, skeletal data has numerous advantages in terms of accuracy and real-time execution when applied to human action recognition. Consequently, many studies have emerged proposing various feature extraction models for skeletal data, among which OpenPose [12] has garnered significant attention and undergone improvements by numerous researchers. For easy institutive, the skeleton data extracted from OpenPose is demonstrated as shown in Figure 1. OpenPose is a model that has been enhanced by the same group of authors from the PAFs (part affinity fields) model [13]. Both models have been implemented on the COCO dataset [14]. The input data for these models is a 2D static image containing multiple objects with different actions, from which the models can identify the objects and extract the skeletal data for each object.



(a) (b) (c) (d)

Fig. 1. Visual represent (a) Input image (b) Part confidence maps (c) Part affinity fields (d) Skeleton data in the OpenPose model

Given an input image of size w × h, the part confidence maps are a set S representing the joint locations of the body parts of the objects. the part affinity fields are a set of vectors used to represent the motion directions of the limbs of the body concerning the joint connections:

$S = (S_1, S_2, S_3, \dots, S_j)$ with j is the number of confidence maps, $S_j \in R^{w \times h}$ (1)

$L = (L_1, L_2, L_3, \dots, L_C)$ with c as the number of vectors, $L_C \in R^{w \times h \times 2}$ (2)

Both OpenPose and PAFs go through multiple stages during their execution. Both models use two branches to predict the estimation sets S and L. However, while PAFs use a kernel of size 7x7, OpenPose utilizes a smaller kernel size of 3x3. Additionally, OpenPose can also detect the skeletal structure of the feet, hands, and even the face of the objects. The process of determining the sets S and L in OpenPose also involves different adjustments compared to the PAFs model. In the study by Yan [15], OpenPose was applied to extract 2D skeletal information from the RGB Kinetics-400 dataset [16], and a convolution graph was used to capture spatial and temporal information.

PoseNet is a model specifically designed for real-time camera re-localization tasks [17]. It employs a regression neural network to learn a mapping from the input image captured by the camera to the position and orientation of the camera in 3D space. This enables the camera to accurately and rapidly

re-localize itself without relying on external sensors such as GPS or markers in the environment. PoseNet's architecture and training process allows it to perform efficient and accurate camera re-localization solely based on visual input. The general flowchart of PoseNet [18] is demonstrated in Figure 2, illustrating how the model operates.

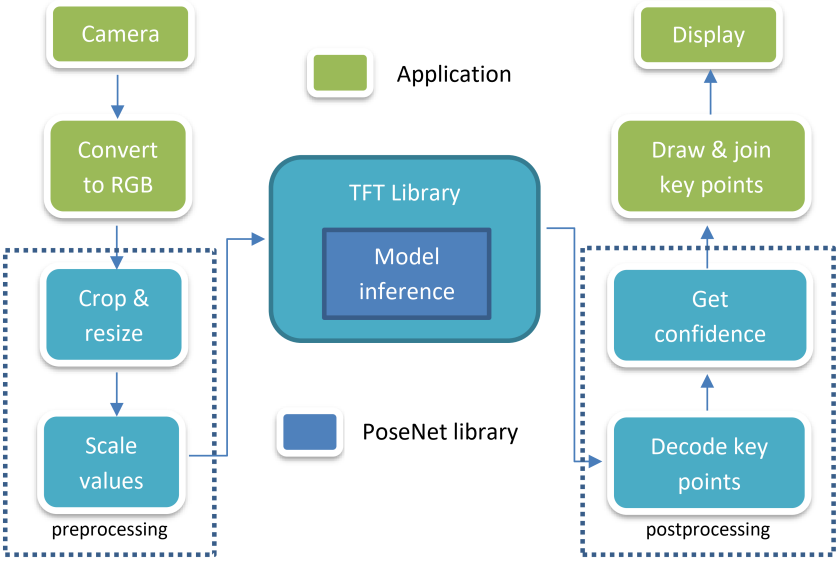


Fig. 2. General flow of PoseNet

In the study [5], the authors utilized both the OpenPose and PoseNet models to construct the MPOSE2021 dataset. This dataset was created by extracting the skeletal data from various subdatasets, including KTH, IXMAS, i3DPost, Weizmann, ISLD, ISLD-AS, UTKinect, and UTD-MHAD. The MPOSE2021 dataset, as depicted in Figure 3, comprises a total of 15,249 samples, each representing one of the 20 different action classes. By leveraging the capabilities of OpenPose and PoseNet, the authors were able to extract and annotate the skeletal data from these diverse subdatasets to create a comprehensive and representative dataset for action recognition research.

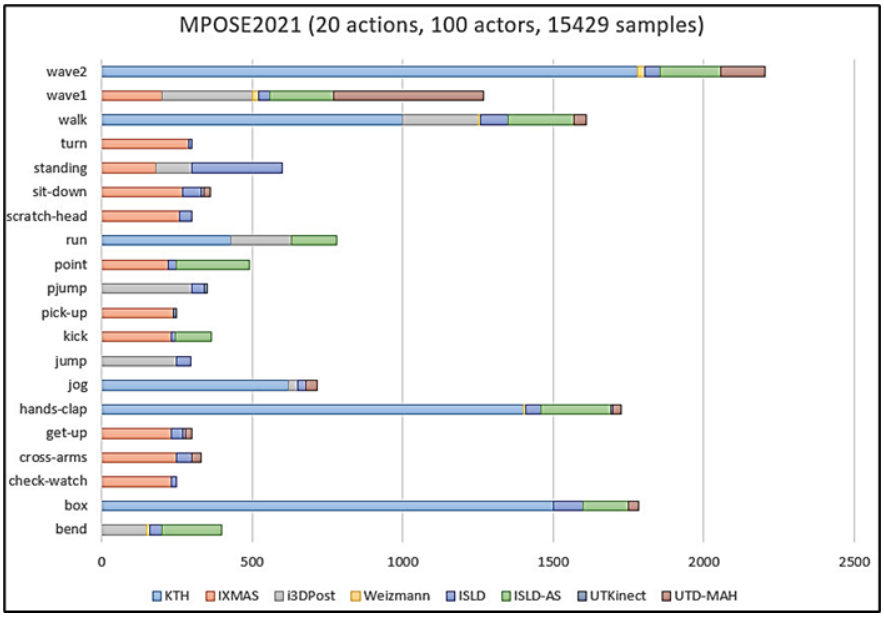


Fig. 3. The number of data samples in MPOSE2021 is divided according to actions and sub-datasets

MSR Action 3D dataset is an important dataset in the field of 3D action recognition. Created by Liu [19] from Microsoft Research, this dataset provides information not only about 2D frame images but also about the 3D spatial information of the actions. MSR Action 3D consists of 567 videos containing 20 different action classes such as walking, jumping, punching, rope jumping, and many others. Each video in the dataset is labeled with its corresponding action. What sets this dataset apart is that it provides 3D spatial information using deeper sensors such as Kinect. The MSR Action 3D dataset has become a valuable resource for the research and development of 3D action recognition methods. Widely used in the research community, this dataset has contributed to improving the performance of 3D action recognition models and algorithms, making significant advancements in this field.

The study [20] introduces DSwarm-Net, a framework that combines deep learning and swarm intelligence-based metaheuristic for Human Action Recognition (HAR) using 3D skeleton data. It extracts four types of features from the skeleton data and encodes them into images: Distance, Distance Velocity, Angle, and Angle Velocity. These encoded images are fed into a modified Convolutional Neural Network (CNN) model. The models are trained, and deep features are extracted from the pre-final layer. The obtained features are optimized using Ant Lion Optimizer to remove non-informative features and reduce dimensionality. DSwarm-Net achieves competitive results on UTD-MHAD, HDM05, and NTU RGB+D 60 datasets, outperforming existing models.

Data augmentation is a widely used technique in image classification to improve performance when labelled data is limited. By enforcing the model's predictions to remain unchanged under diverse data transformations, it introduces desired invariant properties (e.g., lighting invariance) and enhances accuracy. Compared to image data, video data exhibits more complex appearance variations due to the additional temporal dimension. However, data augmentation methods for videos have not been fully explored. In this paper [21], different data augmentation strategies are investigated to capture various invariances in videos, including photometric, geometric, temporal, and actor/scene transformations. When integrated into existing semi-supervised learning frameworks, the authors demonstrate significant improvements in datasets such as Kinetics-100/400, Mini-Something-v2, UCF-101, and HMDB-51 under low-label conditions. Furthermore, the effectiveness of the proposed data augmentation strategy is validated in fully supervised settings, highlighting its ability to enhance performance.

Although skeleton-based action recognition has theoretical advantages in being less affected by environmental factors, in practice, capturing skeleton data depends on the actor's viewpoint and often suffers from errors in joint localization. To address this issue, the research [22] proposes an unsupervised learning method that learns action representations from multiple viewpoints. This approach focuses on maximizing the shared information among different viewpoints of the same action sequence to build a robust representation for human action recognition. Furthermore, the authors introduce a global-local contrastive loss to capture the multi-scale co-occurrence relationships in both spatial and temporal domains. The experimental results illustrate that the proposed approach effectively enhances the performance of unsupervised skeleton-based action recognition on demanding datasets, including PKUMMD, NTU RGB+D 60, and NTU RGB+D 120. These results significantly contribute to the advancement of the field of skeleton-based human action recognition.

In recent years, the graph convolutional network (GCN) has shown great success in extracting features from spatial data, particularly in skeleton-based feature extraction. Nevertheless, the rigid graph structure imposed by the adjacency matrix frequently results in inadequate spatial modeling, subpar generalization, and an excessive number of parameters. In this paper [23], the authors propose a spatially adaptive residual graph convolutional network (SARGCN) for action recognition

based on skeleton features. The proposed method overcomes these issues by allowing flexible graph topology and introducing a learnable parameter matrix, resulting in improved feature extraction and generalization with fewer parameters. Inspired by ResNet, a residual connection is incorporated in the GCN for higher accuracy at lower computational costs. Extensive experiments on NTU RGB+D 60 and NTU RGB+D 120 datasets validate the effectiveness of the proposed approach.

In another related study [24] on GCN, a drawback of using skeleton data in human action recognition is the loss of important cues and related factors, resulting in ambiguous and misclassified actions. To address this issue, the authors proposed an FR Head (Feature Refinement Head) that incorporates spatial-temporal decoupling and contrastive feature refinement. This approach aims to obtain discriminative representations of skeletons and dynamically calibrate ambiguous samples in the feature space. The FR Head can be applied at different stages of GCNs to achieve multi-level refinement and stronger supervision. Extensive experiments conducted on NTU RGB+D, NTU RGB+D 120, and NW-UCLA datasets demonstrated the competitive performance of the proposed models compared to state-of-the-art methods, particularly in discriminating ambiguous samples.

Introducing novel advancements in skeleton-based action recognition, this paper [25] simultaneously addresses three limitations associated with conventional approaches: errors in skeleton detection and tracking, limited variety of targeted actions, and challenges in person-wise and frame-wise action recognition. The authors introduce a point cloud deep-learning paradigm and propose a unified framework with a novel deep neural network architecture called Structured Keypoint Pooling. This approach sparsely aggregates keypoint features based on the inherent structure of skeletons, considering the instances and frames to which each keypoint belongs. It achieves robustness against input errors and expands the range of targeted actions. Additionally, the authors propose a Pooling-Switching Trick inspired by Structured Keypoint Pooling, enabling weakly supervised person-wise and frame-wise action recognition using only video-level action labels. The proposed method demonstrates superior performance compared to state-of-the-art skeleton-based action recognition and spatio-temporal action localization methods. Experimental results validate the effectiveness of the proposed approach in addressing the identified limitations.

Skeleton-based human action recognition has garnered significant attention due to its ability to provide compact and rich high-level representations. However, the challenge of effectively capturing global dependencies among joints during spatio-temporal feature extraction remains. In this paper [26], the Action Capsule method is proposed, which identifies action-related key joints by considering the latent correlation of joints in a skeleton sequence. During the inference stage, the end-to-end network focuses on these key joints, aggregating their spatio-temporal features to recognize the action. The incorporation of multiple stages of action capsules enhances the network's capability to classify similar actions. Comparative analysis with existing methods demonstrates the advantages of the capsule-based approach, particularly in handling missing skeleton data through iterative processing. The proposed network achieves superior performance on the N-UCLA dataset and competitive results on the NTURGBD dataset. Notably, the computational requirements of the authors' approach are significantly reduced based on GFLOPs measurements.

3. Action Transformer V2

In this section, we describe the architecture of the AcTv2 model and provide a summary of some key points of the AcT model and the Transformer architecture.

3.1 Act Model and Transformer Architecture

The AcT (Action Transformer) model takes as input a video sequence consisting of T frames with dimensions H x W x C, denoted as $X_{rgb} \in R^{T\times H\times W\times C}$. Before being fed into the AcT network, the video is pre-processed by a multi-person 2D pose estimation network (F2Dpose) to extract 2D poses. The result of this process is a matrix:

$$X_{2DPose} = F_{2DPose}(X_{rgb}) \tag{3}$$

The AcT model processes each pose sequence X_{2DPose} individually. Initially, the poses in the sequence are projected into a higher-dimensional space (D_{model}) using a linear projection ($W^{l_0} \in R^{P\times D_{model}}$). To generate a general representation for the entire sequence, a class token [CLS] is added at the beginning of the sequence, and a vector of size D_{model} is learned. This class token helps aggregate information from all poses and generates a high-dimensional representation that distinguishes different action classes. Additionally, to provide positional information for the sequence, a positional embedding matrix $X_{pos} \in R^{(T+1)\times D_{model}}$ is added to all tokens to represent their positions in the sequence.

The tokens, including [CLS], are fed into a standard Transformer encoder F_{Enc} with L layers and layer normalization. The result is a matrix $X^L \in R^{(T+1)\times D_{model}}$ that represents the entire encoded sequence. Finally, only the [CLS] token x_{cls}, is passed through a linear classifier head MLP_{Head} to make predictions for the action class.

Meanwhile, the Transformer Encoder utilizes L layers with interleaved self-attention and feed-forward blocks. These blocks are adjusted with Dropout, LayerNorm, and residual connections. The process of aggregating the Encoder blocks is summarized in Figure 4. Each feed-forward block is a multi-layer perceptron with two layers, employing a non-linear activation function. The first layer expands the dimension from D_{model} to D_{mlp} and applies the activation function, while the second layer reduces the dimension back to D_{model}.

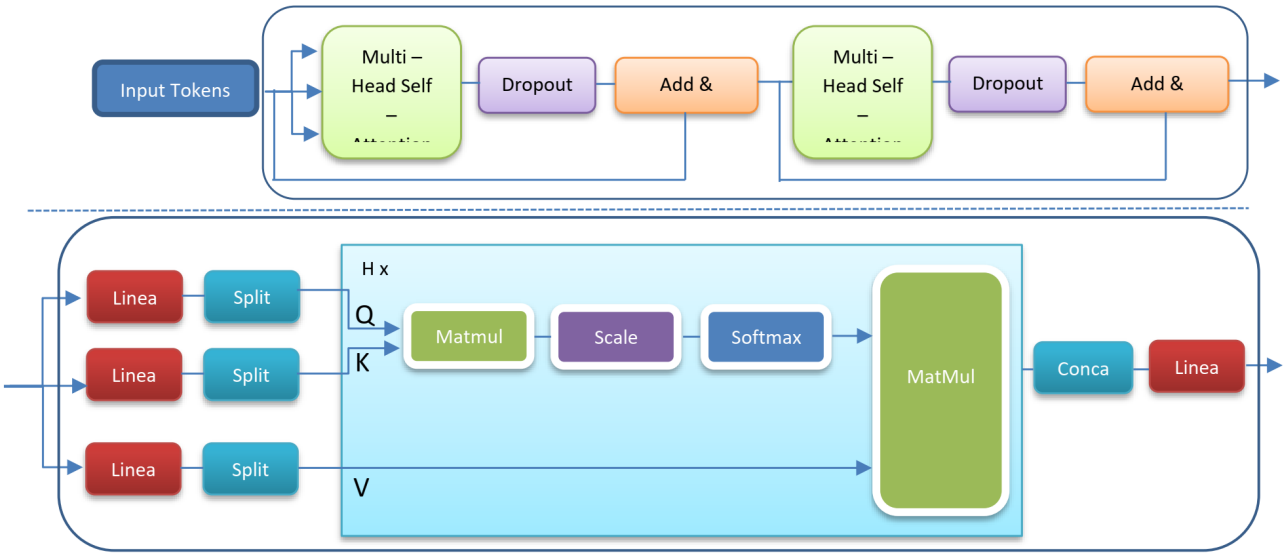


Fig. 4. Transformer architecture in AcT model [5]

3.2 AcTv2 Model

The AcT model has been enhanced and improved by introducing several new layers, such as BatchNormalization1, Dropout, Dense, and BatchNormalization2 before the final Dense output layer of the original model. The BatchNormalization1 layer is used to normalize the input values during the training process, improving the performance and stability of the model. The inclusion of a Dropout layer in the model aids in mitigating overfitting as it introduces randomness by dropping out a fraction of the units during training, thereby reducing the over-reliance on specific features. The Dropout layer operates based on the parameter p = 1 - rate to determine the probability of dropping out a unit, along with the parameter y = x * mask to adjust the output value using the mask = (random tensor < p)/p, which is a matrix with True/False values based on the probability p.

The Dense layer, also known as a fully connected layer, allows the model to establish intricate linear connections between the input and output features, enabling it to learn complex patterns and relationships. It helps capture intricate relationships between input and output features. The BatchNormalization2 layer, placed before the output layer, is used to normalize the output values before feeding them into the output layer. This normalization step improves the stability and prediction capabilities of the model. Figure 5 illustrates the enhanced AcTv2 model based on the AcT architecture. In this depiction, the improvements are showcased through the addition of new layers positioned after the MLP Head block.

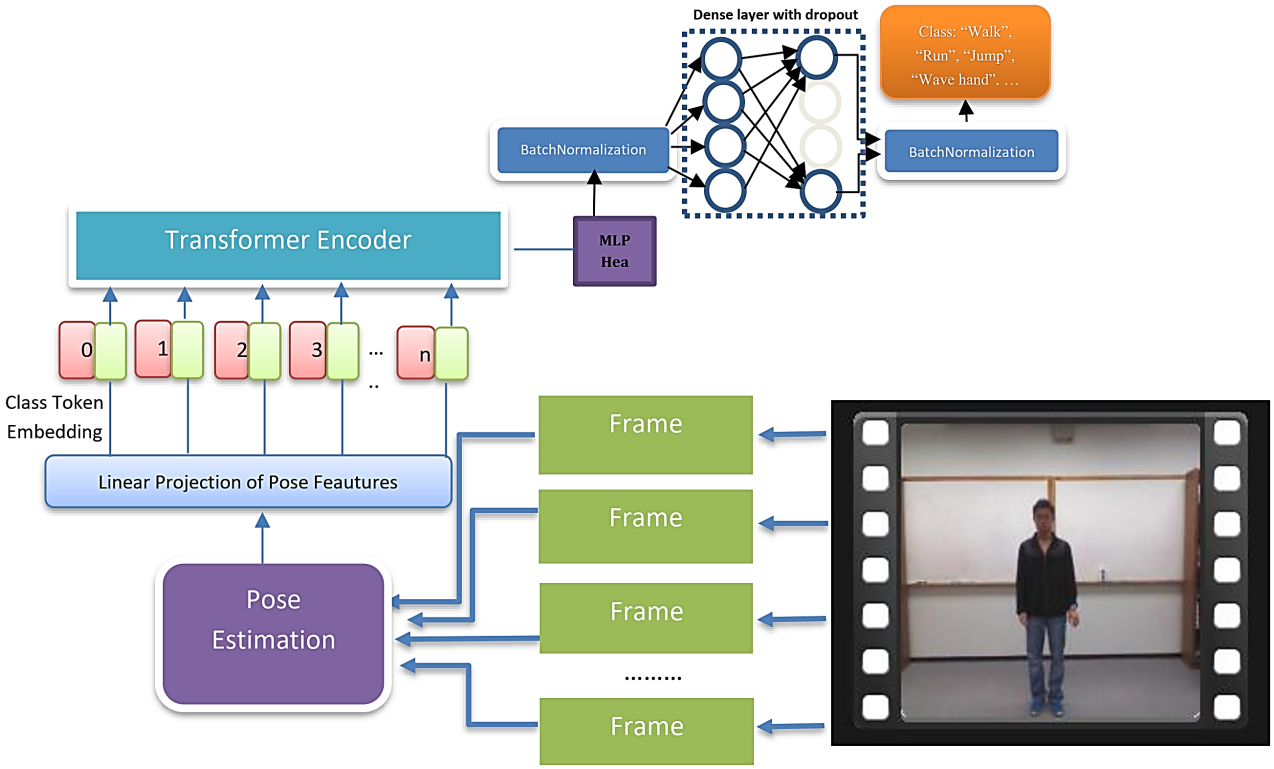


Fig. 5. AcTv2 model architecture with additional layers

Together, these layers contribute to an improved AcT model with better performance and generalization abilities compared to the original AcT model. The specific improvements are as follows. First, we added a BatchNormalization1 layer to normalize the output of the preceding Dense layer of the AcT model. This BatchNormalization1 layer helps adjust and normalize the output values, ensuring that the output values are stable and consistent throughout the model. Next, we applied a Dropout layer to randomly deactivate a portion of the output units during the training process. This

helps to mitigate overfitting and increase the model's generalization ability. The Dropout layer prevents the model from relying too heavily on specific features and ensures that the model can learn more generalized features.

Then, we added a Dense layer to establish fully connected connections between the layers in the model. This Dense layer allows the model to learn complex features and generate non-linear responses, capturing intricate relationships within the input data. Lastly, we added a final BatchNormalization2 layer to normalize the output of the preceding Dense layer before making the final predictions. This ensures that the final prediction results are stable and highly accurate.

Overall, the improvement of the AcT model by incorporating BatchNormalization1, dropout, dense, and batchnormalization2 layers into the original model has significantly enhanced the model's learning capability and generalization ability. These layers have helped the model accurately and reliably recognize complex actions.

4. Action Transformer V2 with MPOSE2021 and MSR Action 3D Datasets

This section presents the experimental results in two scenarios: the experiment of the AcTv2 model on the MPOSE2021 dataset to evaluate the effectiveness of the AcTv2 model, and the experiment of the AcT and AcTv2 models on the MSR Action 3D dataset to evaluate the effectiveness of the two models compared to other solutions.

4.1 Experiment Settings

In addition to retaining some key parameters of the original AcT model, due to the different nature of the two datasets, different configuration parameters were used in the two experimental scenarios, which are presented as shown in Table 1.

Table 1
Hyperparameters are used for experiment processing

Training			
	MPOSE2021 – ACTv2	MSR Action 3D - AcT	MSR Action 3D – AcTV2
Training epochs		7000	7000
Batch size	512	512	512
Optimizer	AdamW	AdamW	AdamW
Warmup epochs	40%	30%	30%
Step epochs	80%	70%	70%
Regularization			
Weight decays	1e-4	1e-4	1e-4
Label smoothing	0.1	0.1	0.1
Dropout-AcT	0.3	0.3	0.3
Dropout-AcTv2	0.5	0.5	0.5
Randomflip	50%	50%	50%
Random noise	0.03	0.03	0.03

We used the TensorFlow framework to train the proposed model on a computer with an Intel i5-13600K CPU and an Nvidia 3090 GPU. Following the mentioned testing strategy, the total training time for the cases was approximately 10 hours. We adhered to the optimization settings and hyperparameters as described in most cases, with adjustments made to the learning rate, number of epochs, and batch size to optimize the training on our dataset and achieve better training results.

4.2 Action Recognition in MPOSE

We conducted thorough experiments on the MPOSE2021 dataset, comparing the AcTv2 model with the AcT model as well as several baseline models and popular HAR architectures presented in [5]. To ensure the accuracy and reliability of the results, we trained 10 different models on different validation splits, keeping a consistent 10% ratio from the training set with a similar class distribution. The experiments were performed on three train/test data splits taken from the MPOSE2021 dataset (Table 2).

Table 2
Benchmark of AcT model and AcTv2 model for short time HAR on MPOSE 2021

	AcT			AcTv2		
	Accuracy	Balance accuracy	Highest accuracy	Accuracy	Balance accuracy	Highest accuracy
Split data1	90.9%	87.1%	91.5%	90.7%	86.7%	92.1%
Split data2	90.9%	84.5%	91.4%	90.9%	84.9%	91.5%
Split data3	90.5%	87.9%	91.6%	90%	87.7%	90.4%

The experimental results show that the AcTv2 model, when trained on the data1 and data2 splits of the MPOSE2021 dataset, can achieve higher accuracy compared to the AcT model. However, this trend does not hold for the data3 split of the dataset.

4.3 Action Recognition in MSR Action 3D

The experiment on the MSR Action 3D dataset involved data preprocessing, model training, and evaluation. Firstly, the data was preprocessed to prepare it in a suitable format for the models. The dataset was split into a training set and a test set, with a 20% ratio for validation data and 80% for training data. Next, the AcT and AcTv2 models were trained on the training data. The training process was carried out with 7000 epochs, and appropriate optimization parameters were set. After the training process, the performance of both models was evaluated on the test set by measuring the accuracy of their predictions.

The evaluation results show that the AcTv2 model has a higher accuracy than the AcT model. Both models outperform previous solutions on the MSR Action 3D dataset in terms of accuracy. This indicates that the AcTv2 model is a significant improvement and has better learning capabilities in action recognition from 3D data.

The experimental results in Table 3 show that the Balance Accuracy metric is significantly better for both the AcT and AcTv2 models. This indicates the stability and effectiveness of the action recognition solution provided by both models when applied to the MSR Action 3D dataset. Furthermore, the AcTv2 model achieves higher accuracy compared to the AcT model, once again confirming the effectiveness of the proposed improvement in this paper.

Table 3
The experimental results on the MSR action 3D dataset

	AcT			AcTv2		
	Accuracy	Balance accuracy	Highest accuracy	Accuracy	Balance accuracy	Highest accuracy
Split data1	88.8%	91.3%	94.2%	89.5%	91.2%	95.7%
Split data2	89.8%	91.2%	94.8%	90%	92.2%	96.5%
Split data3	90.1%	91.4%	94.8%	89.3%	90.1%	95.7%

Based on the findings presented in Table 4, it is evident that the enhanced AcTv2 model achieves remarkable performance gains in action recognition when evaluated on the MSR Action 3D (AS1) dataset. These results highlight the significant advancements made by the proposed model compared to previous research efforts. The superior effectiveness of the AcTv2 model underscores its potential for improving action recognition accuracy and establishing new benchmarks in the field.

Table 4
Accuracy comparison on MSR-Action3D (AS1 sub dataset)

No.	Method	Accuracy
1	Action graph, 2010 [19]	72.9 %
2	Histogram, 2012 [27]	87.98 %
3	Eigen joints, 2012 [28]	74.5 %
4	Conv3DJ, 2013 [29]	88.04 %
5	Joint position (JP), 2014 [30]	93.36 %
6	Relative JP (RJP), 2014 [30]	95.77 %
7	Joint angle (JA), 2014 [30]	84.51 %
8	Absolute SE (3), 2-14 [30]	90.3 %
9	LARP, 2014 [30]	94.72 %
10	Spline curve, 2015 [31]	83.08 %
11	Multi-fused, 2017 [32]	90.8 %
12	CovP3DJ, 2018 [33]	93.48 %
13	ConvMIJ, 2018 [33]	93.48 %
14	Lie algebra with VTDF, 2020 [34]	94.66 %
15	Proposed (AcTv2)	96.5 %

The Table 5 indicates that the AcTv2 model exhibits longer training times than the AcT model on both the Mpose2021 and MSR-Action3D datasets. Specifically, on the Mpose2021 dataset, the training time for the AcTv2 model increased from 8345 seconds to 10023 seconds, while on the MSR-Action3D dataset, the training time increased from 2347 seconds to 13897 seconds. The increase in training time for the AcTv2 model can be attributed to the incorporation of additional layers and features that enhance its performance. While the AcTv2 model demonstrates superior effectiveness in action recognition, it also requires more time to train due to its increased complexity and capacity for capturing more intricate patterns and representations.

Table 5
The training time on OpenPose2021 and MSR-Action 3D

	OpenPose		MSR-action 3D	
	AcT	AcTv2	AcT	AcTv2
Time (s)	8345	10023	2347	13897

The evaluation of training times in the Table 5 also highlights an interesting observation. When the dataset size is relatively small, fine-tuning the parameters to achieve optimal accuracy for the AcTv2 model significantly increases the training time compared to its application on larger datasets. This behaviour is expected as smaller datasets may require more iterations and parameter adjustments to effectively learn from limited samples.

In light of this observation, researchers dealing with smaller datasets face challenges in terms of longer training times and the need for careful parameter tuning. On the other hand, when dealing with larger datasets, the potential increase in training time can be a reasonable trade-off for gaining improved efficiency and performance. The AcTv2 model demonstrates its superiority by achieving

better action recognition results on the evaluated datasets, validating its significance in the field. However, it is crucial for researchers to carefully weigh the trade-off between training time and performance gain based on their specific application requirements and constraints.

5. Conclusion

In this study, we introduced the AcTv2 model as an improvement over the AcT model and evaluated its performance on two distinct datasets: MPOSE2021 and MSR Action 3D. The experimental findings demonstrate the superiority of the AcTv2 model over its predecessor in terms of accuracy and performance.

Specifically, on the MPOSE2021 dataset, the AcTv2 model achieved higher accuracy compared to the AcT model. This indicates that AcTv2 exhibits enhanced action recognition capabilities and delivers reliable results on this dataset. Furthermore, on the MSR Action 3D dataset, the AcTv2 model outperformed not only the AcT model but also 14 other previously published solutions, achieving higher accuracy. The architectural advancements in the AcTv2 model have significantly bolstered its capacity to recognize actions within this 3D dataset, and it stands as one of the top-performing methods. The AcTv2 model is an improvement over the original AcT model by incorporating additional layers and more complex mechanisms to enhance action recognition capabilities. Specifically, AcTv2 utilizes more layers in feature extraction and data processing, thereby strengthening the model's ability to represent information effectively. This leads to higher performance and accuracy in action recognition compared to the AcT model.

However, this enhancement also comes with a trade-off, as the algorithmic complexity of the AcTv2 model has increased. The addition of multiple layers and complex mechanisms results in longer training times and demands more computational resources. Consequently, deploying the AcTv2 model on resource-constrained systems may pose challenges and reduce its practical feasibility and efficiency in real-world environments. Thus, when utilizing the AcTv2 model, careful consideration should be given to the computational resource requirements and training time. Balancing the algorithmic complexity and action recognition performance is crucial. For smaller datasets, the increased complexity may lead to overfitting, causing a decline in performance on new data. To mitigate this, parameter tuning and optimization of the AcTv2 model are necessary to ensure optimal performance on specific datasets.

In conclusion, the transition from the AcT model to the AcTv2 model represents a substantial leap in action recognition performance on both the MPOSE2021 and MSR Action 3D datasets. Despite the increased computational cost associated with its training, the enhanced performance of AcTv2 justifies its adoption in scenarios that demand precise and efficient action recognition. The AcTv2 model holds promise for innovation in various domains, including education and traffic safety. Integrating AcTv2 into attendance management not only brings innovation to education but also opens opportunities for research and technological solutions in monitoring within educational environments [35]. Another potential application of AcTv2 is integration into the field of traffic safety, especially in detecting motorcycle accidents. Research [36] utilized motion sensors on mobile phones to detect accidents and alert emergency services. Implementing AcTv2 could offer more approaches to enhance incident detection and prompt response in emergencies, contributing to improving the safety of motorcycle riders on the road.

Acknowledgement
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Thi, Tuan Hue, Jian Zhang, Li Cheng, Li Wang, and Shinichi Satoh. "Human action recognition and localization in video using structured learning of local space-time features." In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, p. 204-211. IEEE, 2010. https://doi.org/10.1109/AVSS.2010.76

[2] Sun, Zehua, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. "Human action recognition from various data modalities: A review." *IEEE transactions on pattern analysis and machine intelligence* 45, no. 3 (2022): 3200-3225. https://doi.org/10.1109/TPAMI.2022.3183112

[3] Johansson, Gunnar. "Visual perception of biological motion and a model for its analysis." *Perception & psychophysics* 14 (1973): 201-211. https://doi.org/10.3758/BF03212378

[4] Alexey, Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv: 2010.11929* (2020). https://doi.org/10.48550/arXiv.2010.11929

[5] Mazzia, Vittorio, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. "Action transformer: A self-attention model for short-time pose-based human action recognition." *Pattern Recognition* 124 (2022): 108487. https://doi.org/10.1016/j.patcog.2021.108487

[6] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In *31st Conference on Neural Information Processing System (NIPS)* (2017): *1706.03762*. https://doi.org/10.48550/arXiv.1706.03762

[7] Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. "Training data-efficient image transformers & distillation through attention." In *International conference on machine learning*, p. 10347-10357. PMLR, 2021. https://doi.org/10.48550/arXiv.2012.12877

[8] d'Ascoli, Stéphane, Hugo Touvron, Matthew L. Leavitt, Ari S. Morcos, Giulio Biroli, and Levent Sagun. "Convit: Improving vision transformers with soft convolutional inductive biases." In *International conference on machine learning*, p. 2286-2296. PMLR, 2021. https://doi.org/10.48550/arXiv.2103.10697

[9] Yang, Fuzhi, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. "Learning texture transformer network for image super-resolution." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 5791-5800. 2020. https://doi.org/10.48550/arXiv.2006.04139

[10] Zhu, Hongyu, Hao Liu, Congcong Zhu, Zongyong Deng, and Xuehong Sun. "Learning spatial-temporal deformable networks for unconstrained face alignment and tracking in videos." *Pattern Recognition* 107 (2020): 107354. https://doi.org/10.1016/j.patcog.2020.107354

[11] Dong, Linhao, Shuang Xu, and Bo Xu. "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition." In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5884-5888. IEEE, 2018. https://doi.org/10.1109/ICASSP.2018.8462506

[12] Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, no. 1 (2021): 172-186. https://doi.org/10.1109/TPAMI.2019.2929257.

[13] Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Realtime multi-person 2d pose estimation using part affinity fields." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 7291-7299. 2017. https://doi.org/10.1109/TPAMI.2019.2929257

[14] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740-755. Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-10602-1_48

[15] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1. 2018. https://doi.org/10.1609/aaai.v32i1.12328

[16] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299-6308. 2018. https://doi.org/10.48550/arXiv.1705.07750

[17] Li, Ming, Jiangying Qin, Deren Li, Ruizhi Chen, Xuan Liao, and Bingxuan Guo. "VNLSTM-PoseNet: A novel deep ConvNet for real-time 6-DOF camera relocalization in urban streets." *Geo-spatial Information Science* 24, no. 3 (2021): 422-437. https://doi.org/10.1080/10095020.2021.1960779

[18] Jo, BeomJun, and SeongKi Kim. "Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices." *Traitement du Signal* 39, no. 1 (2022): 119. https://doi.org/10.18280/ts.390111

[19] Li, Wanqing, Zhengyou Zhang, and Zicheng Liu. "Action recognition based on a bag of 3d points." In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, p. 9-14. IEEE, 2010. https://doi.org/10.1109/CVPRW.2010.5543273

[20] Basak, Hritam, Rohit Kundu, Pawan Kumar Singh, Muhammad Fazal Ijaz, Marcin Woźniak, and Ram Sarkar. "A union of deep learning and swarm-based optimization for 3D human action recognition." *Scientific Reports* 12, no. 1 (2022): 5494. https://doi.org/10.1038/s41598-022-09293-8

[21] Zou, Yuliang, Jinwoo Choi, Qitong Wang, and Jia-Bin Huang. "Learning representational invariances for data-efficient action recognition." *Computer Vision and Image Understanding* 227 (2023): 103597. https://doi.org/10.1016/j.cviu.2022.103597

[22] Bian, Cunling, Wei Feng, Fanbo Meng, and Song Wang. "Global–local contrastive multiview representation learning for skeleton-based action recognition." *Computer Vision and Image Understanding* 229 (2023): 103655. https://doi.org/10.1016/j.cviu.2023.103655

[23] Zhu, Qilin, and Hongmin Deng. "Spatial adaptive graph convolutional network for skeleton-based action recognition." *Applied Intelligence* 53, no. 14 (2023): 17796-17808. https://doi.org/10.1007/s10489-022-04442-y

[24] Hu, Lizhang, and Jinhua Xu. "Learning discriminative representation for skeletal action recognition using LSTM networks." In *Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part II 17*, pp. 94-104. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-64698-5_9

[25] Hachiuma, Ryo, Fumiaki Sato, and Taiki Sekii. "Unified keypoint-based action recognition framework via structured keypoint pooling." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 22962-22971. 2023. https://doi.org/10.1109/CVPR52729.2023.02199

[26] Bavil, Ali Farajzadeh, Hamed Damirchi, and Hamid D. Taghirad. "Action capsules: Human skeleton action recognition." *Computer Vision and Image Understanding* 233 (2023): 103722. https://doi.org/10.1016/j.cviu.2023.103722

[27] Xia, Lu, Chia-Chih Chen, and Jake K. Aggarwal. "View invariant human action recognition using histograms of 3d joints." In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20-27. IEEE, 2012. https://doi.org/10.1109/CVPRW.2012.6239233

[28] Yang, Xiaodong, and Ying Li Tian. "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor." In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 14-19. IEEE, 2012. https://doi.org/10.1109/CVPRW.2012.6239232

[29] Hussein, Mohamed E., Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban. "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations." In *Twenty-third international joint conference on artificial intelligence*. 2013. https://dl.acm.org/doi/10.5555/2540128.2540483

[30] Vemulapalli, Raviteja, Felipe Arrate, and Rama Chellappa. "Human action recognition by representing 3d skeletons as points in a lie group." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 588-595. 2014. https://doi.org/10.1109/CVPR.2014.82

[31] Ghorbel, Enjie, Rémi Boutteau, Jacques Boonaert, Xavier Savatier, and Stéphane Lecoeuche. "3D real-time human action recognition using a spline interpolation approach." In *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 61-66. IEEE, 2015. https://doi.org/10.1109/IPTA.2015.7367097

[32] Jalal, Ahmad, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, and Daijin Kim. "Robust human activity recognition from depth video using spatiotemporal multi-fused features." *Pattern recognition* 61 (2017): 295-308. https://doi.org/10.1016/j.patcog.2016.08.003

[33] El-Ghaish, Hany A., Amin A. Shoukry, and Mohamed E. Hussein. "CovP3DJ: Skeleton-parts-based-covariance descriptor for human action recognition." In *VISIGRAPP (5: VISAPP)*, pp. 343-350. 2018. https://doi.org/10.5220/0006625703430350

[34] Boujebli, Malek, Hassen Drira, Makram Mestiri, and Imed Riadh Farah. "Rate-invariant modeling in lie algebra for activity recognition." *Electronics* 9, no. 11 (2020): 1888. https://doi.org/10.3390/electronics9111888

[35] Rosman, Mohamad Rahimi Mohamad, Mohamad Iqmal Ussaiq Ismail, Muhamad Hakim Ahmad Dzarawi, and Muhamad Alif Zhafri Md Azman. "Conceptualizing an attendance monitoring system for Malaysian Educational Institutions." *Journal of Advanced Research in Computing and Applications* 16, no. 1 (2019): 10-23.

[36] Ghazalli, Hajar Izzati Mohd, Muhammad Izaiddin Hassan, Zuhri Arafah Zulkifli, and Siti Nuramalina Johari. "MotoSOS: Accident detection for motorcycle riders using motion sensors." *Journal of Advanced Research in Computing and Applications* 15, no. 1 (2019): 9-19.