View-invariant 3D Skeleton-based Human Activity Recognition based on Transformer and Spatio-temporal Features

Ahmed Snoun, Tahani Bouchrika and Olfa Jemai
Research Team in Intelligent Machines (RTIM),
National Engineering School of Gabes (ENIG), University of Gabes, Gabes, Tunisia

Keywords: Human Activity Recognition, 3D Skeleton, Spatio-temporal Features, View-invariant, Transformer.

Abstract: With the emergence of depth sensors, real-time 3D human skeleton estimation have become easier to accomplish. Thus, methods for human activity recognition (HAR) based on 3D skeleton have become increasingly accessible. In this paper, we introduce a new approach for human activity recognition using 3D skeletal data. Our approach generates a set of spatio-temporal and view-invariant features from the skeleton joints. Then, the extracted features are analyzed using a typical Transformer encoder in order to recognize the activity. In fact, Transformers, which are based on self-attention mechanism, have been successful in many domains in the last few years, which makes them suitable for HAR. The proposed approach shows promising performance on different well-known datasets that provide 3D skeleton data, namely, KARD, Florence 3D, UTKinect Action 3D and MSR Action 3D.

1 INTRODUCTION

Many computer vision applications, such as intelligent surveillance, human-computer interaction, and robotics, rely on Human Action Recognition or HAR for short. Despite tremendous improvements, precisely predicting what humans do in unseen videos remains a difficult task due to a variety of difficulties, such as viewpoint changes, intra-class variation, and environment distractions. Currently, depth sensor-based HAR is regarded as one of the most promising approaches for solving the aforementioned challenges. Depth sensors that are inexpensive can provide 3D structural information of the human body, which is proven useful for the HAR task. Most of these sensors, in particular, have real-time skeleton estimation algorithms (Shotton et al., 2013) that are resistant to distractions in the environment. As a result, using skeletal data for HAR opens up possibilities for overcoming RGB and depth modalities' constraints. Therefore, there are two major concerns that need to be addressed for skeleton-based action recognition. The first problem is figuring out how to turn raw skeletal sequences into a useful representation that can capture the spatio-temporal dynamics of human motions. The second step is to use the motion representation acquired from skeletons to model and recognize actions. HAR based on hand-crafted features and HAR based on deep learning models are the two primary groupings of previous works on this topic. The first set of approaches extracts hand-crafted local information from skeletal joints and employs probabilistic graphical models to represent and categorize actions, such as the Hidden Markov Model (HMM) (Lv and Nevatia, 2006), Conditional Random Field (CRF) (Han et al., 2010), and Fourier Temporal Pyramid (FTP) (Vemulapalli et al., 2014). Many approaches for skeleton-based action recognition (Xia et al., 2012)(Vemulapalli et al., 2014)(Wang et al., 2014)(Wu and Shao, 2014)(Wang et al., 2016) have been presented since the first study on 3D HAR using depth data (Li et al., 2010). These approaches all have one thing in common: they extract geometric features from the 3D coordinates of the skeletal joints and use a generative model to describe their temporal information. Despite the promising results of these approaches, most of them can be easily affected by the change of Kinect's viewpoint. In order to meet with the view variance problem, we propose in this paper a view-invariant approach based on the extraction of spatio-temporal geometric features from 3D skeleton data after changing the reference point from the Kinect center to one of the skeleton's keypoints.

The second group treats skeleton-based action recognition as a time-series problem and recommends that the temporal evolutions of skeletons be analyzed using Recurrent Neural Networks with Long Short-Term Memory units (RNN-LSTMs). For that mat-

Snoun, A., Bouchrika, T. and Jemai, O.
View-invariant 3D Skeleton-based Human Activity Recognition based on Transformer and Spatio-temporal Features.
DOI: 10.5220/0010895300003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 706-715
ISBN: 978-989-758-549-4; ISSN: 2184-4313

ter, many works used RNN-LSTMs to model the activity and have obtained good results (Du et al., 2015)(Zhu et al., 2016)(Liu et al., 2016)(Liu et al., 2017)(Peng et al., 2021). Despite the good performance of RNN-LSTMs, recent studies tend to change them with Transformer (Vaswani et al., 2017), which has been one of the most important deep learning introductions for natural language processing (NLP) in recent years. In addition to NLP, self-attention mechanism has been shown to be effective for a variety of tasks, including image classification (Dosovitskiy et al., 2020), generative adversarial networks (Lin et al., 2018), and speech recognition (Berg et al., 2021), which proves that this architecture is equally appropriate for HAR. In this paper, we use a pure Transformer encoder architecture to analyze the view-invariant spatio-temporal features extracted from 3D skeletal data.

This work's primary contributions can be summarized as follows:

- **Transformation of the coordinate system origin from the kinect center to the head and spine:** we propose a view-invariant approach for HAR by changing the reference point of the Kinect camera to the spine and head of the human skeleton and thus the coordinates of the remaining keypoints will be relative to the spine and the head of the skeleton and not the camera and therefore will not be affected by the change of the kinect's view.
- **Generation of spatio-temporal features:** we extract spatio-temporal features from the 3D skeletal data. The spatial features are the 3D coordinates of the joints relative to both the head and the spine and the distances between the keypoints and the head and spine. However, the temporal features are the movement angles as well as the movement type of each joint between each two frames.
- **Analyzing the extracted features with Transformer encoder:** we introduce a new HAR system based on the transformer encoder and we demonstrate that fully self-attention architectures outperform other models in the HAR task.

The remainder of this paper will be as follows: Section 2 is dedicated to review the works of the literature. We explain in section 3 the different aspects of the proposed approach. Section 4 is reserved to present and discuss the experimental results. Finally, we conclude and propose future works in section 5.

2 RELATED WORKS

In recent decades, researchers have looked into several compact representations of human activity. Johansson's experiment from 1975 shown that individuals can perceive activity with very little observers (Johansson, 1975). A sequence of a person walking in a dark room with lights mounted to the individual's major joints was used by Johansson to explain his point. Despite the fact that just light specks were seen, the 3D motion in these clips was clearly discernible. Therefore, various works in the literature widely used skeletal data to recognize human activities. These methodologies are known as skeleton-based HAR. The input to these systems is made up of a series of points that represent body joints. Skeletons can be represented in the 2D space (Jlidi et al., 2020)(Snoun et al., 2021) of the image view or 3D space, such as those generated with Kinetic sensors. In this work, we focus on 3D pose information since it gives a better representation of the human pose and motion.

Devanne et al. (Devanne et al., 2013) suggested spatio-temporal motion trajectories to model human actions. The distance between two curves is then represented using an elastic metric, which is a metric that is invariant to the speed and time of the action within a Riemannian form space. Finally, using a k-Nearest-Neighbor (k-NN) classifier, the action recognition task was viewed as a classification in Riemannian space. In (Gan and Chen, 2013), the relative locations and local spherical angles of the APJ3D representation are computed using a selection of 15 skeletal joints. Following the key postures selection, they used an updated Fourier Temporal Pyramid (Wang et al., 2012) and random forests to classify the action. Another technique for modeling joints is HOJ3D (Xia et al., 2012), which divides the 3D space into n bins and uses a Gaussian weight function to connect the joints with each bin. Then, using a clustering approach, a discrete Hidden Markov Model (HMM) is used to model the postures' evolution in time. A human activity can also be described by a mixture of static skeleton features, which reflect the current frame, successive motion features, which are computed using the current and prior frames, and overall dynamics features, which take into account the current and previous frames (Yang and Tian, 2014). In (Kim and Kim, 2015), a view-invariant method for HAR was proposed, it is based on pose estimation using 3D body pose stream. To model the activities, the authors extracted motion, structure and hand positions from joints coordinates. To make their approach view-invariant, the generated 2D spherical co-

ordinates based on polar angle and azimuthal angle, and they calculated relative positions of left and right hand by head and torso. The extracted features were analyzed with a Hidden-state Conditional Random Field (HCRF) to recognize the activity. The joint's spherical coordinates were used also by Taha et al. (Taha et al., 2015) to describe the skeleton. Then, a multiclass SVM and a discrete HMM were used to distinguish activities made up of several actions. Others used also a combined machine learning algorithm to classify actions, such as Gaglio et al. (Gaglio et al., 2015), who used a multiclass SVM to model postures and a discrete HMM to describe an action as a succession of poses. Human activities are also viewed as a temporal flow of body poses (Theodorakopoulos et al., 2013), and skeletal data is processed to create invariant pose representations, which are represented by eight pairs of angles.

Deep learning and specifically RNN and LSTM were widely used in the literature in the aim of HAR. For example, an end-to-end RNN with handmade subnets was proposed by Du et al. (Du et al., 2015). The raw locations of body joints are separated into five pieces and fed into five bidirectional RNNs according to human structure. The network hierarchically merged the representations derived by the subnets to a higher-level representation as the layers number increased. Zhu et al. (Zhu et al., 2013) used LSTM to analyze the skeletal sequence. The LSTM input consists of the 3D concatenated locations of skeletal joints in a frame for each step. They used a set of encoded feature vectors to model a feature manifold. Finally, the manifold was used to help and regularize LSTM supervised learning for action recognition using RGB video. A 2D Spatio-Temporal LSTM framework was developed in (Liu et al., 2018) to investigate the hidden sources of activity-related context information in both the temporal and geographical domains. They also suggested a trust gate method that would deal with the depth sensors' imprecise 3D joint coordinates. More recently, Noori et al. (Noori et al., 2019) proposed a HAR approach based on motion features extracted from skeletal data. Therefore, they extracted the magnitude and angle of each joint in the human body. To classify the extracted features, they used an RNN-LSTM based network.

Unlike earlier techniques, we introduce, in this paper, an architecture for HAR that is based purely on the Transformer encoder (Vaswani et al., 2017), with no convolutional or recurrent layers. To train the transformer, we used spatio-temporal features extracted from 3D joints coordinates of the human skeleton. To deal with the view variance problem, we generated new joints coordinates relative to the head and spine of the human body instead of the camera position. Hence, the activity recognition process will not be affected by the change of the camera view.

3 METHODOLOGY

The human body is an articulated system of rigid segments connected by joints, and human action is thought to be a continuous evolution of the spatial and temporal configuration of these segments (skeletons). Therefore, we design a system based on spatio-temporal features extracted from the body joints to recognize human activity. To emphasize, as illustrated in Figure 1, the designed system starts with 3D skeletons as input. Then, we extract spatio-temporal geometric features from the skeletal data to better model the spatial and temporal flow of the human body that represents the activity. The extracted features are analyzed later with a Transformer encoder in order to output the activity label.

3.1 Spatio-temporal Features Generation

This step consists in generating a vector of view-invariant spatio-temporal geometric features from each skeleton of the input sequence. Therefore, Microsoft's Kinect API directly provides a set of 3D joint locations, which may be calculated from depth images acquired from the Kinect sensor. The 3D position (x, y, z) of each joint provided by the Kinect API, on the other hand, is represented using the Cartesian coordinate system, with the origin (0, 0, 0) in the Kinect sensor's center. As a result, if either the Kinect sensor or the target object moves, the 3D position data of a joint can be simply altered. This means that the 3D joint coordinates obtained straight from the Kinect API are extremely sensitive to Kinect's view variation, and hence are ineffective characteristics for reliably identifying daily human activities in a variety of environments. The problem of view variance is depicted in Figure 2. If the position of Kinect is changed, the corresponding position value of the same joint collected by the Kinect sensor will not be the same ($P_1 = (x_1, y_1, z_1) \neq P_2 = (x_2, y_2, z_2)$).

To prevent the view variance problem, we change the origin of the coordinate system form the center of the Kinect sensor to the spine and head of the skeleton, as shown in Figure 3. Therefore, the positions of the different joints will be relative to the spine and head and no longer to the Kinect sensor. Thus, the change of this latter's position will not affect anymore the reliability of the skeletons coordinates. Here, we

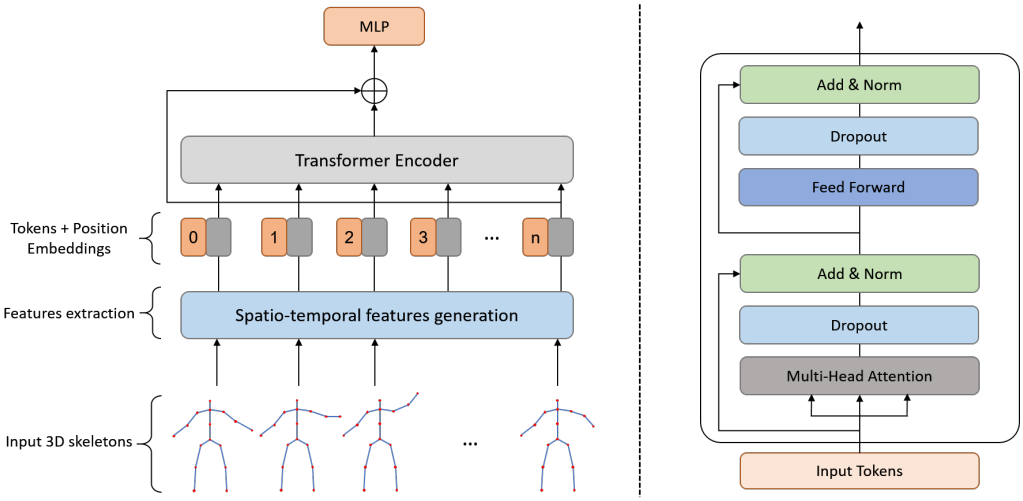

Figure 1: Overview of the HAR architecture based on spatio-temporal geometric features and Transformer Encoder (left) and Transformer encoder layer architecture (right).

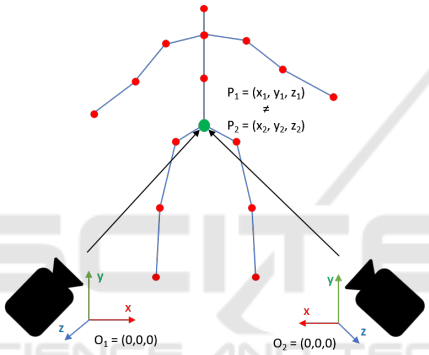

Figure 2: Illustration of the view variance problem.

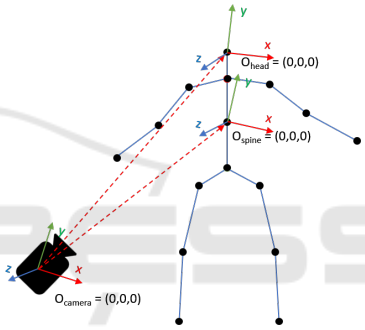

Figure 3: Transformation of the coordinate system origin from the kinect sensor to the head and spine.

chose the head and spine as the new centers of the coordinate system because they have less motion compared to the hands and legs and by consequence, they are less involved in the human motion. To translate the coordinate system origin to the head and spine, we multiplied each joint position (x_j, y_j, z_j) by the translation matrix as represented in equations 1 and 2.

$$\begin{pmatrix} \hat{x_j} \\ \hat{y_j} \\ \hat{z_j} \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -x_{head} \\ 0 & 1 & 0 & -y_{head} \\ 0 & 0 & 1 & -z_{head} \\ 0 & 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} x_j \\ y_j \\ z_j \\ 1 \end{pmatrix} \quad (1)$$

$$\begin{pmatrix} \hat{x_j} \\ \hat{y_j} \\ \hat{z_j} \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -x_{spine} \\ 0 & 1 & 0 & -y_{spine} \\ 0 & 0 & 1 & -z_{spine} \\ 0 & 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} x_j \\ y_j \\ z_j \\ 1 \end{pmatrix} \quad (2)$$

Once we obtain the set of joints coordinates relative to head and spine, we compute a set of spatial and temporal features.

3.1.1 Spatial Features

The best way to model the spatial structure of the skeleton is to calculate the distance between each joint of the skeleton and the head in a first time and the spine in a second time as illustrated in Figure 4. Therefore, we calculate the Euclidean distance between each joint j of each frame t and the head

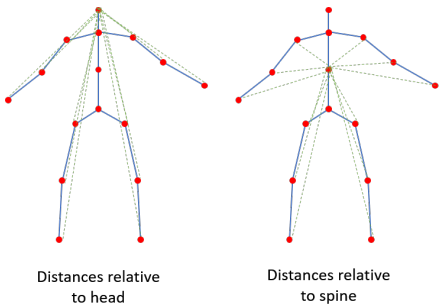

Figure 4: Spatial features generation. (left) Distances between each joint and the head. (right) distances between each joint and the spine.

(D_{jt_head}) and the distance between each joint j of each frame t and the spine (D_{jt_spine}) using respectively equations 3 and 4:

$$D_{jt\_head} = \sqrt{(x_{jt} - x_{head})^2 + (y_{jt} - y_{head})^2 + (z_{jt} - z_{head})^2} \quad (3)$$

$$D_{jt\_spine} = \sqrt{(x_{jt} - x_{spine})^2 + (y_{jt} - y_{spine})^2 + (z_{jt} - z_{spine})^2} \quad (4)$$

3.1.2 Temporal Features

The temporal features of a joint j in a frame t are obtained by computing the movement angles and movement direction compared to the same joint in the previous frame $t-1$, as shown in Figure 5.

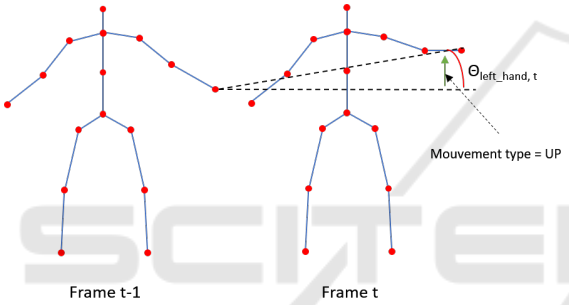

Figure 5: Temporal features generation.

Let $j_t(x_{jt}, y_{jt}, z_{jt})$ be the position of joint j in frame t and $j_{t-1}(x_{jt-1}, y_{jt-1}, z_{jt-1})$ be the position of joint j in frame $t-1$. The movement angle of a joint j in a frame t (θ_{jt}) is calculated using the following equation:

$$\theta_{jt} = \cos^{-1}\left(\frac{\vec{Oj_t} \cdot \vec{Oj_{t-1}}}{\|Oj_t\| \cdot \|Oj_{t-1}\|}\right) \quad (5)$$

Where $O = (0,0,0)$ is the origin of the coordinate system (head or spine), $\|Oj_t\| = \sqrt{x_{jt}^2 + y_{jt}^2 + z_{jt}^2}$ and $\|Oj_{t-1}\| = \sqrt{x_{jt-1}^2 + y_{jt-1}^2 + z_{jt-1}^2}$

The movement direction is a combination of three possible directions (Right/left, up/down and forward/backward), for example, a joint j in a frame t can move right/down/forward. To find the movement direction of a joint j in a frame t, we compare the coordinates (x_{jt}, y_{jt}, z_{jt}) of joint j in a frame t with the coordinates (x_{jt-1}, y_{jt-1}, z_{jt-1}) of the joint j in the frame $t-1$. If $x_{jt} > x_{jt-1}$ then the movement is right, else if $x_{jt} < x_{jt-1}$ then the movement is left, otherwise there is no movement. The same is applied for y to determine if the movement is up or down and z for forward and backward. Finally, we encoded the movement direction as follows: Right (1), left (2), up (1), down (2), forward (1), backward (2) and no movement (0). For instance, if the movement direction is right/down/forward, the code will be 121.

Once we compute the spatial and temporal features, the input vector to the Transformer encoder will be composed of the transformed 3D coordinates of the joints relative to the head and spine, the distances between each joint and the head as well as the spine, the movements angles between each two frames of each joint and the movement direction of each joint between each two frames.

3.2 Transformer Encoder Architecture

In (Vaswani et al., 2017), Transformer networks were first used for machine translation. The encoder and decoder are the two main elements of the transformer network. An input sequence (source) is received by the encoder, which is processed by a stack of identical layers, including a multi-head self-attention layer and a fully-connected feed-forward network. The encoder's representation is then used by a decoder to construct an output sequence (target). For classification purpose, it is recommended to use only the Transformer encoder. Therefore, to classify our input sequence, we use a typical encoder architecture. As illustrated in the right of Figure 1, the transformer encoder layer E is composed of multi-head self-attention and feed-forward blocks. After each block, Dropout, LayerNorm, and residual connections are applied. Each feed-forward block is a two-layer perceptron (ff) with GeLu (Hendrycks and Gimpel, 2016) as non-linearity. In the proposed implementation, the first layer uses non-linearity and expands the dimension from D_{model} to $D_{mlp} = 2 * D_{model}$. The second layer, on the other hand, reduces the dimension from D_{mlp} to D_{model} (D_{model} is the length of the input features vector).

By passing the input via a LayerNorm (LN) before each module and putting it back with residual connections, each layer E in the Transformer Encoder performs the following computation:

$$\hat{E}(S) = LN(S + Dropout(MHA(S))) \quad (6)$$

$$E(S) = LN(\hat{E}(S) + Dropout(ff(\hat{E}(S)))) \quad (7)$$

Where $ff(X) = Linear(Dropout(GeLu(Linear(X))))$ denotes the feed-forward block explained above, MHA is the Multi-head self-attention block and S in the input sequence. In our case, S is a set of spatio-temporal features vectors extracted from the skeleton joint as explained in the previous section,

summed with a positional embedding matrix that represents the positionality information.

The transformer network's self-attention mechanism is a critical component. The architecture of the self-attention mechanism block is shown in Figure 6. Therefore, a function that expresses a weighted sum of the values V is called attention A. The weights are determined by comparing a query Q to a collection of keys K. The scaling dot-product is the most used form of the matching function. Attention with the scaled dot-product matching function (A), which is illustrated in the right of Figure 6, is written in formal terms as:

$$A(Q,V,K) = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V \quad (8)$$

Where d denotes the dimension of both queries and keys.

The Multi-head attention MHA is an extension of attention that uses independent linear projections h_i of (Q,K,V) to create numerous parallel attention functions:

$$MHA(Q,V,K) = \text{Concat}(h_1,...,h_m)W \quad (9)$$

$$h_i = A(QW_i^Q, KW_i^K, VW_i^V) \quad (10)$$

Finally, the output of the transformer encoder $E(S)$ is fed into a linear classification head (MLP_{head}) after applying a residual connection in order to allow the gradients to flow through the network directly and enhance the performance. The computation of the MLP_{head} is as follows:

$$\hat{y} = MLP_{head}(S + E(S)) \quad (11)$$

Where $\hat{y}$ are the output logits, which are passed later to a Softmax activation function in order to find the final activity.

4 EXPERIMENTAL RESULTS

4.1 Used Datasets

The **KARD dataset** (Gaglio et al., 2015) contains 18 activities divided into 10 gestures and eight actions. This dataset was collected in an office environment with a Kinect camera placed 2-3 meters away from the person. The activities were carried out by 10 people (nine males and one female). Each person repeated each activity three times, resulting in 540 sequences. The dataset provides 15 skeleton joints in world and screen coordinates.

The **Florence3D dataset** (Seidenari et al., 2013) includes 9 different activities: waving, drinking from a bottle, answering the phone, clapping, tightening lace, sitting, standing up, reading a watch, and bowing. These activities were carried out by 10 different subjects twice or three times, for a total of 215 sequences. The activities were captured in a variety of settings, and only RGB videos and 15 skeleton joints are available.

The **UTKinect dataset** (Xia et al., 2012) consists of ten different subjects (nine males and one female) performing ten different activities twice. The dataset includes the following activities: walking, sitting down, standing up, picking up, carrying, throwing, pushing, pulling, waving, and clapping hands. Since there is one unlabeled sequence, a total of 199 sequences are available, with sample actions ranging in length from 5 to 120 frames. The dataset contains 20 skeleton joints captured with the Kinect sensor with a rate of 15 fps.

Finally, one of the most commonly used datasets for HAR is the **MSR Action 3D** (Li et al., 2010). It consists of 20 activities performed twice or three times by 10 subjects. There are 567 skeleton frame sequences in total; each skeleton is composed of 20 joints. High arm waving, horizontal arm waving, hammering, hand catching, forward punching, high throwing, drawing X, drawing tick, drawing circle, hand clapping, two hand waving, side boxing, bending, forward kicking, side kicking, jogging, tennis swinging, tennis serving, golf swinging, and picking up and throwing are all activities included in the dataset. The data was collected at 15 fps with a structured-light depth camera.

4.2 Obtained Results

To validate our approach, we followed the evaluation protocol used by previous works for each of the above-mentioned datasets.

4.2.1 KARD Dataset

Three different experiments on five different activities groups of the dataset were proposed by the collectors of the KARD dataset (Gaglio et al., 2015). The experiments are as follows:

- Experiment A: Train/Test split: 30/70%
- Experiment B: Train/Test split: 70/30%
- Experiment C: Train/Test split: 50/50%

The dataset's activities are divided into the following categories:

- **Gestures:** horizontal arm waving, high arm waving, two hand waving, high throwing, drawing x, drawing tick, forward kicking, side kicking, bending, hand clapping

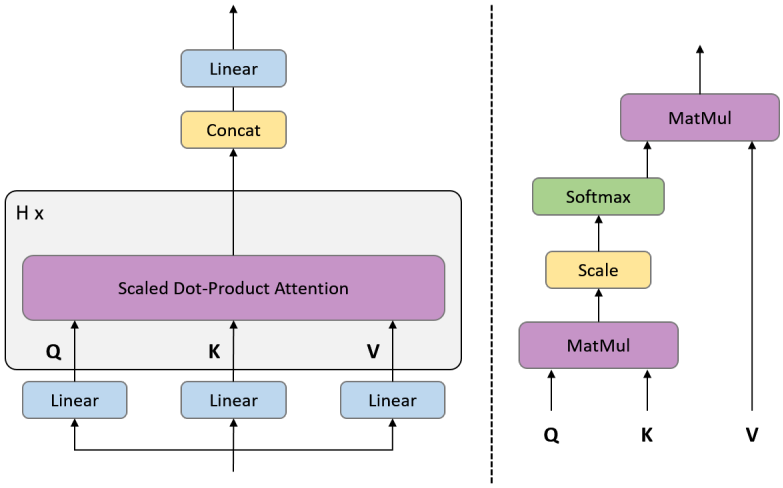

Figure 6: Multi-head self-attention block (left) and Scaled dot-product attention block (right).

- **Actions:** catching cap, tossing paper, taking umbrella, walking, phone calling, drinking, sitting down, standing up
- **Activity Set 1:** horizontal arm waving, two hand waving, bending, phone calling, standing up, forward kicking, drawing x, walking
- **Activity Set 2:** high arm waving, side kicking, catching cap, drawing tick, hand clapping, forward kicking, bending, sitting down
- **Activity Set 3:** drawing tick, drinking, sitting down, phone calling, taking umbrella, tossing paper, high throwing, horizontal arm waving

Each group was tested three times following the three evaluation strategies mentioned earlier. The results of Gestures and Actions categories are reported in Table 1. According to the results, we notice that for Gestures, the evaluation protocol B, where the 2/3 of the data is used to train the model and the rest to test it, outperforms the two other protocols. In fact, Experiment B outperforms the two other experiments in all of the groups (see Table 1 and Table 2), which means giving more data for training helps to improve the recognition performance. We notice also that the Actions set seems to be easier to analyze since the results of the three experiments outperform the three experiments in the Gestures group. This can be explained by the fact that in the Actions set most of the body joints are engaged in the activity, unlike the Gestures where only few joints are engaged.

Table 2 summarizes the results of the Activity Set tests. It is shown that Activity Set 1 and 2, which seem to be the simplest ones, have good recognition results compared to Activity Set 3. The reported results show also that the proposed approach outperforms the original work proposed in (Gaglio et al., 2015), which is based on the use of a discrete HMM to describe an activity as a succession of poses and a multiclass SVM to classify poses. It outperforms also the work in (Cippitelli et al., 2016), which is based on the extraction of key poses to create a feature vector and classification using a multiclass SVM.

Finally, the “new-person” evaluation protocol is also performed. This scenario consists in using nine of the ten persons of the dataset to train the model and using the remaining person to test. Therefore, since there is no recommendation on how split the dataset, we used the 18 activities to perform the “new-person” scenario. The obtained results are reported in Table 3. We notice that our model outperforms the works in (Gaglio et al., 2015) and (Cippitelli et al., 2016) by respectively 11% and 1% in terms of precision and recall.

4.2.2 Florence 3D Dataset

The leave-one-actor-out setting, which is equivalent to the previously described ”new-person” setting, is used to evaluate this dataset. The obtained results using the proposed transformer approach as well as a comparison with the works of the state-of-the-art are drawn in Table 4. The proposed approach achieved an accuracy equals to 91%. We notice that our accuracy is less than the accuracy (96.2%) reported in (Taha et al., 2015). However, the proposed approach outperforms the other mentioned literature works, such as the work in (Seidenari et al., 2013), which uses a bag-of-poses to model the skeleton data, and the work in (Vemulapalli et al., 2014) which uses Lie Group to model the human action as a curve.

In fact, the Florence 3D dataset is challenging due to the high interclass similarity, which means the existence of similar activities like drink from a bottle and

Table 1: Accuracy (%) of the proposed approach on the KARD dataset with the dataset split in Gestures and Actions using different experiments.

	Gestures			**Actions**		
	A	B	C	A	B	C
(Gaglio et al., 2015)	86.5	93.0	86.7	92.5	95.0	90.1
(Cippitelli et al., 2016)	89.9	95.9	93.7	99.1	99.9	99.4
Transformer (ours)	96.7	100	97.3	98.8	100	100

Table 2: Accuracy (%) of the proposed approach on the KARD dataset with different Activity Sets using different experiments.

	Activity Set 1			**Activity Set 2**			**Activity Set 3**		
	A	B	C	A	B	C	A	B	C
(Gaglio et al., 2015)	95.1	99.1	93.0	89.0	94.9	90.1	84.2	89.5	81.7
(Cippitelli et al., 2016)	98.2	99.0	98.1	99.8	100	99.7	91.6	95.8	93.3
Transformer (ours)	98.8	100	99.17	100	100	100	94.5	98.6	95.0

Table 3: Precision (%) and recall (%) of the proposed approach on the KARD dataset using the "new-person" setting.

	Precision	**Recall**
(Gaglio et al., 2015)	84.8	84.5
(Cippitelli et al., 2016)	95.1	95.0
Transformer (ours)	96.7	96.3

Table 4: Accuracy (%) of the proposed approach on the Florence 3D dataset using the "new-person" setting.

	Accuracy
(Seidenari et al., 2013)	82.0
(Vemulapalli et al., 2014)	90.9
(Taha et al., 2015)	96.2
(Cippitelli et al., 2016)	86.1
Transformer (ours)	91.0

answer phone, and also the high intraclass similarity which means that an action can be performed in different ways by the same person like using once the left hand and once the right hand to perform the same activity.

4.2.3 UTKinect Action 3D Dataset

As in the case of the Florence 3D dataset, the work in (Xia et al., 2012) proposed the leave-one-out cross-validation (LOOCV) evaluation protocol, which is similar to the "new-person" protocol. Table 5 shows the obtained results on this dataset compared to previous works. The results show that the work in (Vemulapalli et al., 2014) produced the best accuracy, but their approach is very complex since they model the skeleton in a Lie group and then, before classification with one-versus-all multiclass SVM, they process it with Dynamic Time Wrapping to achieve temporal alignments and a special representation known as the Fourier Temporal Pyramid. In the other hand, our approach outperforms the original work on this dataset (Xia et al., 2012) with 3.1% in term of accuracy. The key performance constraint with this dataset is the limited number of frames that some sequences contain compared to other sequences, which can decrease the results dramatically.

Table 5: Accuracy (%) of the proposed approach on the UTKinect Action 3D dataset using the LOOCV setting.

	Accuracy
(Xia et al., 2012)	90.9
(Theodorakopoulos et al., 2013)	90.95
(Zhu et al., 2013)	91.9
(Gan and Chen, 2013)	92.0
(Vemulapalli et al., 2014)	97.1
(Cippitelli et al., 2016)	95.1
(Liu et al., 2017)	92.0
Transformer (ours)	94.0

4.2.4 MSR Action 3D Dataset

In the literature, there is a lot of uncertainty on which validation parameters should be used with MSR Action 3D dataset. Padilla-Lopez et al. (Padilla-López et al., 2014) suggested employing all feasible combinations of 5-5 subject splitting, which consists in using 252 combinations of 5 persons for training and 5 persons for testing, or using leave-one-actor-out (LOAO) protocol, which is similar to the "new-person" scenario. Here, we considered the "new-person" protocol since it is used by most of the works in the literature. According to the results reported in Table 6, we were able to reach an accuracy of 93%, which is better than all of the previous works. Therefore, the work in (Azary and Savakis, 2013) proposed an approach based on skeleton data. However, Chaaraoui et al. (Chaaraoui et al., 2013) used

more advanced techniques, such as the fusion of depth and skeleton data, or selection of the best set of joints. More recent works exploits subspace clustering and temporal pruning to recognize activity (Paoletti et al., 2021). The work in (Zhao et al., 2019) proposed a bayesian hierarchical dynamic model for action recognition. Finally, a Graph based skeleton model is used in (Kao et al., 2019) to recognize the human activity.

Table 6: Accuracy (%) of the proposed approach on the MSR Action 3D dataset using the LOAO setting.

	Accuracy
(Azary and Savakis, 2013)	78.5
(Chaaraoui et al., 2013)	90.6
(Cippitelli et al., 2016)	81.2
(Zhao et al., 2019)	86.1
(Kao et al., 2019)	74
(Paoletti et al., 2021)	88.51
Transformer (ours)	93.0

5 CONCLUSION

In this paper, a view-invariant HAR approach based on 3D skeleton data has been proposed. A spatio-temporal features generation step has been implemented. Therefore, after translating the origin of the skeleton's coordinate system from the center of the camera to the head and spine in order to prevent the view variance problem, we compute the distances between each joint and the new origins and the movement angles as well as the movement direction of each joint between each two consecutive frames of the input sequence. The extracted features are then analyzed by a pure Transformer Encoder in order to recognize the activity associated to these features. Our approach shows improvements compared to most of the state-of-the-art approaches after being tested on the KARD, Florence 3D, UTKinect Action 3D and MSR Action 3D datasets. In future works, we aim to use more sophisticated methods like Graph Neural Networks since the human skeleton can be easily modeled as a graph and evaluate bigger and more challenging datasets such as NTU-RGB-D dataset.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of this work by grants from General Direction of scientific Research (DGRST), Tunisia, under the ARUB program.

REFERENCES

Azary, S. and Savakis, A. (2013). Grassmannian sparse representations and motion depth surfaces for 3d action recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 492–499.

Berg, A., O'Connor, M., and Cruz, M. T. (2021). Keyword transformer: A self-attention model for keyword spotting. *ArXiv*, abs/2104.00769.

Chaaraoui, A. A., Padilla-López, J. R., and Flórez-Revuelta, F. (2013). Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. *2013 IEEE International Conference on Computer Vision Workshops*, pages 91–97.

Cippitelli, E., Gasparrini, S., Gambi, E., and Spinsante, S. (2016). A human activity recognition system using skeleton data from rgbd sensors. *Computational Intelligence and Neuroscience*, 2016.

Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., and Del Bimbo, A. (2013). Space-time pose representation for 3d human action recognition. page 456–464.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Du, Y., Wang, W., and Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118.

Gaglio, S., Re, G. L., and Morana, M. (2015). Human activity recognition process using 3-d posture data. *IEEE Transactions on Human-Machine Systems*, 45:586–597.

Gan, L. and Chen, F.-T. (2013). Human action recognition using apj3d and random forests. *J. Softw.*, 8:2238–2245.

Han, L., Wu, X., Liang, W., Hou, G., and Jia, Y. (2010). Discriminative human action recognition in the learned hierarchical manifold space. *Image Vision Comput.*, 28(5):836–849.

Hendrycks, D. and Gimpel, K. (2016). Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Jlidi, N., Snoun, A., Bouchrika, T., Jemai, O., and Zaied, M. (2020). PTLHAR: PoseNet and transfer learning for human activities recognition based on body articulations. In *12th International Conference on Machine Vision (ICMV 2019)*, volume 11433, pages 187 – 194.

Johansson, G. (1975). Visual motion perception. *Scientific American*, 232 6:76–88.

Kao, J.-Y., Ortega, A., Tian, D., Mansour, H., and Vetro, A. (2019). Graph based skeleton modeling for human activity analysis. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2025–2029.

Kim, H. and Kim, I. (2015). Human activity recognition as

time-series analysis. *Mathematical Problems in Engineering*, 2015:1–9.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14.

Lin, C.-H., Yumer, E., Wang, O., Shechtman, E., and Lucey, S. (2018). St-gan: Spatial transformer generative adversarial networks for image compositing. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9455–9464.

Liu, J., Shahroudy, A., Xu, D., Kot, A. C., and Wang, G. (2018). Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:3007–3021.

Liu, J., Shahroudy, A., Xu, D., and Wang, G. (2016). Spatio-temporal lstm with trust gates for 3d human action recognition. *ArXiv*, abs/1607.07043.

Liu, J., Wang, G., Hu, P., Duan, L.-Y., and Kot, A. C. (2017). Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In Leonardis, A., Bischof, H., and Pinz, A., editors, *Computer Vision – ECCV 2006*, pages 359–372, Berlin, Heidelberg. Springer Berlin Heidelberg.

Noori, F. M., Wallace, B., Uddin, M. Z., and Tørresen, J. (2019). A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network. In *SCIA*.

Padilla-López, J. R., Chaaraoui, A. A., and Flórez-Revuelta, F. (2014). A discussion on the validation tests employed to compare human action recognition methods using the msr action3d dataset. *ArXiv*, abs/1407.7390.

Paoletti, G., Cavazza, J., Çiğdem Beyan, and Bue, A. D. (2021). Subspace clustering for action recognition with covariance representations and temporal pruning. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6035–6042.

Peng, W., Shi, J., Varanka, T., and Zhao, G. (2021). Rethinking the st-gcns for 3d skeleton-based human action recognition. *Neurocomputing*, 454:45–53.

Seidenari, L., Varano, V., Berretti, S., Bimbo, A., and Pala, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–485.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124.

Snoun, A., Jlidi, N., Bouchrika, T., Jemai, O., and Zaied, M. (2021). Towards a deep human activity recognition approach based on video to image transformation with skeleton data. *Multimedia Tools and Applications*, 80:29675–29698.

Taha, A., Zayed, H. H., Khalifa, M. E., and El-Horbaty, E.-S. M. (2015). Human activity recognition for surveillance applications. In *ICIT 2015*.

Theodorakopoulos, I., Kastaniotis, D., Economou, G., and Fotopoulos, S. (2013). Pose-based human action recognition via sparse representation in dissimilarity space. *Journal of Visual Communication and Image Representation*, 25:12–23.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *ArXiv*, abs/1706.03762.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297.

Wang, P., Li, W., Ogunbona, P., Gao, Z., and ling Zhang, H. (2014). Mining mid-level features for action recognition based on effective skeleton representation. *2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8.

Wang, P., Yuan, C., Hu, W., Li, B., and Zhang, Y. (2016). Graph based skeleton motion representation and similarity measurement for action recognition. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 370–385, Cham. Springer International Publishing.

Wu, D. and Shao, L. (2014). Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27.

Yang, X. and Tian, Y. (2014). Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1):2–11. Visual Understanding and Applications with RGB-D Cameras.

Zhao, R., Xu, W., Su, H., and Ji, Q. (2019). Bayesian hierarchical dynamic model for human action recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7725–7734.

Zhu, W., Lan, C., Xing, J., Li, Y., Shen, L., Zeng, W., and Xie, X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. page 8.

Zhu, Y., Chen, W., and Guo, G. (2013). Fusing spatiotemporal features and joints for 3d action recognition. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 486–491.