STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **12**, May 2024, pp 782–798.
Published online in International Academic Press (www.IAPress.org)

Self-Attention Network Assisted by Object Knowledge Integration for Skeleton-Based Action Recognition

Oumaima Moutik*, Hiba Sekkat, Taha Ait tchakoucht, Badr El kari, Ahmed El Hilali Alaoui

Euromed Research Center School of Digital Engineering and Artificial Intelligence Euromed University of Fes, Morocco

Abstract 3D-Skeleton-based action recognition has been widely adopted due to its efficiency and robustness to complex backgrounds. While it is capable of conveying a significant amount of information regarding the dynamics of human poses, we argue that its performance is curtailed when confronted with actions involving interactions between humans and objects due to the absence of the study of the surrounding objects. It is of great importance to delve deeper into the study of human-object interactions for skeleton-based action recognition. This paper proposes a novel approach to represent the spatial-temporal skeleton features, along with the present nearby objects and their dynamics. To accomplish this, a new formulation named object knowledge is presented, which entails the categorization of object characteristics, based on whether or not the object necessitates a motion analysis. With a piece of prior knowledge, in cases where it is required, the motion is calculated, while for those where it is not necessary, only the category of object is considered. This object knowledge is then early-fusion along with the skeleton representation, in such a way that it fits into the self-attention model. The experimental results on different popular action recognition datasets (NTU RGB+D 60, NTU RGB+ D 120) illustrate that the proposed approach outperforms the current state-of-the-art methods.

Keywords Skeleton-based Action Recognition, Human-object Interactions, Self-attention Network, Spatial-Temporal Video Analysis

DOI: 10.19139/soic-2310-5070-1967

1. Introduction

Computer Vision entails teaching computers to interpret and comprehend visual information from images and videos, in different domains of applications such as medical [1], industrial [2], and material [3]. Action recognition has undergone significant advancements to meet the needs of a diverse array of industrial applications, such as video surveillance [4], autonomous driving [5], and robotics [6]. In numerous industrial contexts, the ability to recognize human actions holds the potential to yield significantly valuable information, enabling the early detection of dangerous situations [7], analysis of work behaviors [8], and evaluation of labor productivity [9]. Therefore, there has been a shift in focus among computer vision (CV) researchers towards Skeleton-based action recognition, which entails the organization of actions based on the spatial relationship of body joints as a set of 3D joint coordinates, and the temporal movement of human posture; However, the primary challenge lies in acquiring discriminative and robust spatiotemporal characteristics that can accurately depict different actions. To address this challenge, researchers have applied deep learning models, such as Convolutional Neural Networks (CNNs) [10], Recurrent Neural Networks (RNNs)[11], Graph Convolutional Neural Networks (GCNs)[12], and more recently, Self-attention mechanisms (Self-Att) or vision transformers (ViT) [13].

Skeleton data offers a high-level semantic representation of human action sequences that is more resilient to variations in appearance, viewpoints, and surrounding environments with a low-dimensional representation

*Correspondence to: Oumaima Moutik (o.moutik@ueuromed.org). Euromed Research Center School of Digital Engineering and Artificial Intelligence, Euromed University of Fes, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

compared to RGB data [14]. However, the skeleton representation provides a more abstract representation of the human body, which may result in the loss of some fine-grained details and many actions may be challenging to represent using only skeleton data. The aforementioned observations evolve a motivation for proposing a fusion of Skeleton and RGB data, which capitalizes on the unique advantages of each modality and addresses their respective limitations. Therefore, RGB data can serve as a complementary cue for analyzing the surrounding environment. Nevertheless, the dense features of RGB lie in the object details. Hence, it is clever to disregard extraneous information such as the background and focus solely on studying the objects surrounding the human subject. Particularly, in the presence of proficient pre-trained RGB object detection models and pioneering datasets that encompass most of the requisite objects while skeleton data remain focused on human behavior, ultimately discerning potential actions. Certainly, there exist studies upon the same principle; however, this remains an unsettled and relatively unexplored issue, and in general cases, researchers handle the operation through several different training times, leading to amplified parameter numbers and heightened computational expenses [15], [16].

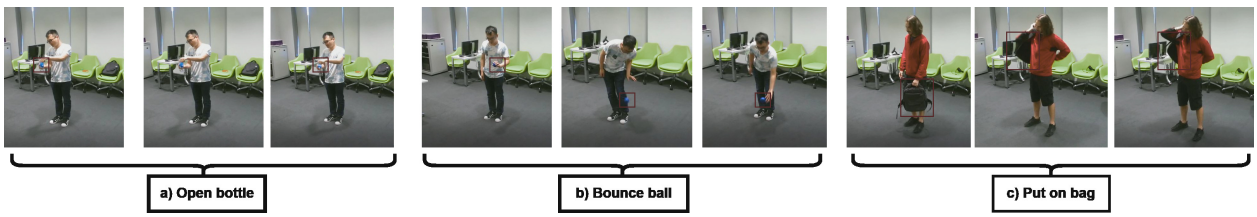

Figure 1. The figure highlights the significance of object knowledge and especially motion in accurately recognizing actions. ”Open bottle”, ”bounce ball” and ”put a bag” are three examples that showcase this importance. More than that, the existence of an object in the scene is not enough to distinguish between actions? nature. For instance, in scenario a), the bag is present in the scene, but the subject does not interact with it, rendering it irrelevant to the action. On the other hand, in scenario c), the subject is putting on the bag, and the object's motion is directly related to the subject's movement, thus making it crucial to identify the action of putting on the bag. This illustrates the critical role of considering an object's motion in action recognition

The present paper introduces an effective approach for action recognition, through skeleton-based analysis, augmented by object-level knowledge obtained from pre-trained RGB data, in which a novel encoding technique is proposed to handle both skeleton features and object knowledge as a unified input for the training network. Object knowledge encompasses both the position and dynamic features of detected objects. The former pertains to non-movable, interactive objects, while the latter refers to movable interactive objects within a given scene. This partition proved to be a more precise identification of interactive objects with the subject as illustrated in Figure 1.

The main contribution of this work can be briefly summarized as follows: (i) A designed set of object knowledge of interactive non-movable objects and interactive movable objects that aligns with the representation of 3D skeleton data to enhance recognition accuracy. (ii) The proposition of a novel fusion technique that processes 3D skeleton data in conjunction with object knowledge. (iii) Based on this, the design of a spatiotemporal network achieves state-of-the-art performance. The approach is evaluated on three extensively utilized datasets, NTU RGB+D 60 [17], NTU RGB+D 120 [18], and PKU-MMD dataset [19] for skeleton-based action recognition.

The subsequent sections of this paper are organized as follows: Section II provides a review of the relevant literature pertaining to the proposed method. Section III presents a detailed description of how the action recognition task is improved through the integration of object knowledge with the self-attention mechanism. The experimental results and their interpretations are discussed in Section IV. Lastly, the conclusion is presented in Section V.

2. Related work

In this section, we succinctly examine relevant literature that is closely related to the method proposed.

2.1. 3D Skeleton-based Action Recognition

The Microsoft Kinect sensor is the source of 3D skeleton data, which captures the human skeleton from depth images and tracks 25 joints of up to 6 individuals in real-time. Each skeleton's joints are represented as (x, y, z) and are connected to form a matrix. The processing of this data is categorized into two main categories: traditional and deep learning methods. Traditional methods manually extract skeletal features using a series of 3D operations, such as rotation angle, translation, and velocity of skeletal joints [20]. Deep learning methods are classified as RNN-based, CNN-based, or GCN-based. RNN methods consist of a recurrent layer that extracts temporal information between joints as a sequence of vectors [11]. CNN methods convert skeleton data into pseudo-image and then use CNN to learn skeleton features [21], [22]. However, neither CNN nor RNN has been able to accurately represent the structure of the skeleton data, as it is a non-Euclidean space and is therefore naturally embedded as a graph, rather than as a sequence of vectors or a 2D grid. This has motivated researchers to develop a more appropriate type of skeleton modeling - a GCN that specifically captures skeleton data as a graph structure [12, 23, 24, 25, 26, 27]. In recent times, Self-Attention (Self-Att) has gained attention among researchers for its performance in capturing long-range dependencies and relationships within spatiotemporal data [28, 29, 30, 30, 31]. Typically, Self-Att is incorporated into the architecture of GCNs or CNNs as a complementary mechanism [32], [33]. However, it should be noted that these methods solely focus on skeleton representation and do not consider additional information from the surrounding environment as a potential source of data. In our paper, we aim to leverage the contextual information provided by action-related objects to enhance the understanding of actions in videos.

2.2. Skeleton-based Human-Object-Interaction

The main goal of Human-Object Interaction (HOI) detection in videos is to identify and extract "human, object, and interaction" triplets. However, existing approaches mainly rely on RGB-based methods, overlooking the potential of skeleton-based modality. As a result, only a limited number of researchers have explored this modality for fine-grained analysis of human-centric videos through HOI analysis [16], [15]. Previous studies have explored the use of RGB modality for object position and skeleton modality for human position analysis. However, the challenge of effectively handling double modality persists. To tackle this issue, we propose a novel method that optimally leverages both modalities through early fusion. Specifically, we combine RGB-dense features with the skeleton modality to form the input representation of the self-attention model. Notably, some prior works, such as Xu et al. [16], did not account for the temporal dynamics of objects in human-object interactions and treated all detected objects equally. Wang et al. [15] introduced a multi-stream network with three streams, which were later fused by averaging the classification scores of each stream and not trained together. Our approach distinguishes itself by employing a one-stream network that solely focuses on object features alongside skeleton modality features. This design choice enables us to better address the challenges associated with double modality while maintaining a simpler and more effective architecture.

2.3. Self-Attention Mechanism

Deriving Inspiration from the way humans selectively focus on specific aspects of a visual scene to better comprehend information [34], attention mechanisms automatically identify and highlight prominent regions or features within input images or feature maps. Self-attention, a type of attention mechanism, emphasizes interdependencies within data [35]. This is achieved through the computation of three trainable weight matrices - query (Q), key (K), and value (V), each with a dimension of d. Subsequently, the dot product of the query and key is normalized by $\sqrt{d}$ to stabilize the gradients, and the resultant product is multiplied by the value to produce the output. In essence, the entire process can be succinctly represented as follows in Equation (1):

$$Attention(Q, K, V) = Z = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

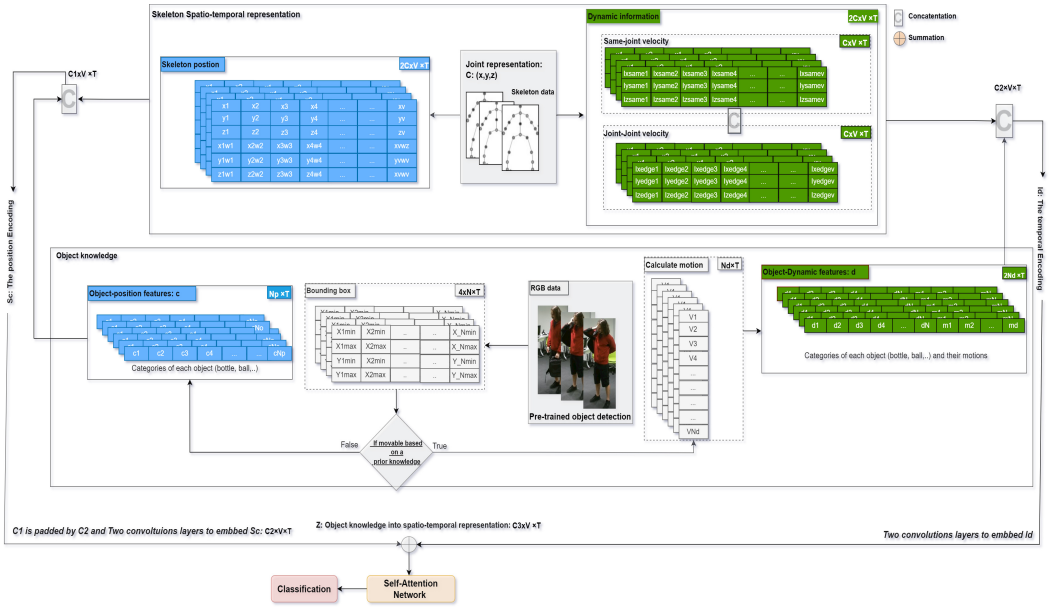

Figure 2. The proposed approach comprises two primary components: position encoding and dynamics encoding. Initially, RGB data is fed into a pre-trained model to detect objects. Subsequently, prior knowledge is utilized to determine if the detected object requires motion analysis. If not, the Object-position encoding is followed by combining the object-position features (c) with the skeleton position. Conversely, if the object necessitates a motion study, the dynamics encoding is applied which is based on including the object category and its motion (d) as Object Dynamic features with the dynamic features of the spatiotemporal representation of the skeleton

3. Our method

3.1. Overall Architecture

Most approaches in the state of the art represent the skeleton data, i.e., the joints, as matrix $P \in R^{C\times V\times T}$ where C denotes the channel number, V and T denote the number of joints and frames. Authors previously enriched the skeleton representation with bone information and joint type as a semantic guide for the neural network [36]. The bone information indicates the direction of joints and joint type mentions the type of joint (head, feet...). Accordingly, as presented in Figure 2, our goal is to further lead and assist the 3D skeleton representation to lessen action recognition misconstrue, by incorporating the elements from the milieu. For this purpose, we commence with the establishment of an object knowledge aggregate that accumulates the position and dynamic information separately of encountered skeleton-related objects. position information provides insights into non-movable objects, while dynamic information investigates also the rate of objects' motion of movable objects. This information is embedded within the 3D skeleton representation as an input of a self-attention architecture as an early fusion.

3.2. A formulation of Object Knowledge data

To collect and focus selectively on the surrounding elements that are relevant for the task of action recognition, this section proposes an aggregate, which we named Object Knowledge, which contains two levels: Object-position Features and Object-dynamic Features.

3.2.1. Object-position Features

Given an RGB video X with T frames, we employed an object detector model on the T frames to identify and detect present elements. The model is named End-to-End Object Detection with Transformers (DE-TR) [37] since it outperforms other existing methods in different terms such as accuracy, and speed, and is known to perform well in detecting small objects. It returns a set of bounding box coordinates, typically represented as a tuple of four values $Bbox = (x_{min}, y_{min}, x_{max}, y_{max})$ indicating the coordinates of the top-left and bottom-right corners of the bounding box, and class labels for each object detected in each T frames, represented as integers, with each integer corresponding to a specific class (e.g. person, car, dog, etc.). The model may also output a confidence score for each bounding box. We used the Object356 dataset [38] to train the DE-TR model. Although it is well known that the COCO dataset [39] is the one that is most frequently used for object detection tasks, Object365 covers a larger number of richer interactive objects about 365 as opposed to the 80 in COCO. We select the N_p first interactive non-movable objects candidates by detection scores. To avoid size inconsistencies, we put $N_p = V$ and for less than N_p we fill with a placeholder of zero. For X, the bounding box of each object forms a matrix in $Bbox \in R^{4\times N_p \times T}$. However, our concern is the $c \in R^{N_p \times T}$ represents the categories of objects.

3.2.2. Object-dynamic Features

Because standalone position information does not provide all the knowledge required to determine which object is interacting with the human in a scenario with several objects, what results in, objects with a significant movement deserve more attention than others. To this end, we exploit the motion of each movable object in each frame. Firstly, we filtered the interacted-movable objects N_d (The same thing as the position representation: we put $N_p = V = N_d$ to avoid dimensional size inconsistencies), and we performed the calculation of motion as a Euclidean distance task of each object N_d between adjacent frames (t) and $(t+1)$ as follows (Equation (2)):

$$V^{t+1} = \| \, c(x,y)^{t+1} - c(x,y)^{t} \, \| \tag{2}$$

where c^t refers to the centroid of the object at frame t, as seen in Figure 3. $c(x)^t = x^t_{min} + \frac{x^t_{max}}{2}$, and $c(y)^t = y^t_{min} + \frac{y^t_{max}}{2}$.

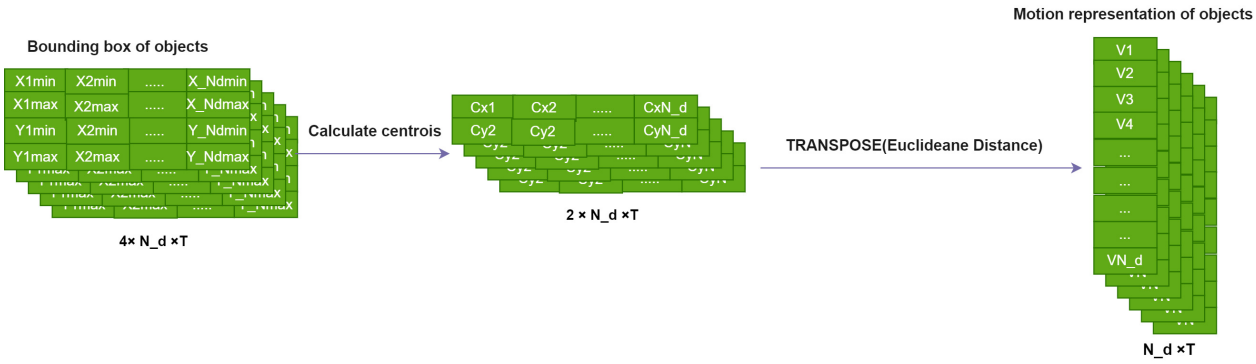

Figure 3. Extraction of the Object-dynamic features of a moving object

At this point, we put d as the dynamics representation of interactive movable objects, it covers the temporal aspect of the objects in addition to its positional representation, distributed in $R^{2N_d \times T}$.

3.3. Object knowledge along with the 3D skeleton representation

3.3.1. 3D Skeleton features encoding

The original skeleton data of position is $P \in R^{C\times V\times T}$. P is enriched with the bone representation. Every joint is converted to a vector that points from the previous joint to the present one, with the root joint's vector remaining at zero, a W_{jj} identity matrix is initially provided, and some elements with the same directed connection joints' column indices are set to be set to -1 to facilitate computation. $E = P \cdot W$ is the information extracted from bones. E and P are then concatenated as: $S = [P \quad P.W]$, where $S \in R^{2C\times V\times T}$.

The temporal information of the skeleton is divided into two elements: Same-Joint velocity and Joint-to-Joint velocity. The same-joint velocity is the velocity of the same joint between adjacent frames. The joint-to-joint velocity is the velocity of the bone made up of the joint j and $j+1$. Given a Skeleton P consisting of T frames. Let p_j^t and p_j^{t-1} be the $3D$ coordinates of the j_{th} joint of the skeleton P at the frame t and $t-1$ respectively. The same-joint velocity of p_j^t is calculated in the form:
On the plane $z = 0$:

$$d_{p_j}^{t,z=0} = \| \ p(x,y)_j^t - p(x,y)_j^{t-1} \ \|, \tag{3}$$

On the plane x = 0:

$$d_{p_j}^{t,x=0} = \| \ p(y,z)_j^t - p(y,z)_j^{t-1} \ \|, \tag{4}$$

On the plane y = 0:

$$d_{p_j}^{t,y=0} = \| \ p(z,x)_j^t - p(z,x)_j^{t-1} \ \| \ . \tag{5}$$

The same-joint velocity of p_j^t is formed from (Equation (3), Equation(4), and Equation(5)) as $v_j^t = [d_{p_j}^{t,z=0}, d_{p_j}^{t,x=0}, d_{p_j}^{t,y=0}] \in R^C$.

Let $e_{ij}^t = p_i^t - p_j^t$ with i and j are two adjacent joints be the bone representation at the frame t. Similar to the same-joint velocity v_i^t, The joint-to-joint velocity is $v_{e_{ij}^t} = [d_{ij}^{t,z=0}, d_{ij}^{t,x=0}, d_{ij}^{t,y=0}]$ with $v_{e_{ij}^t} \in R^C$.

all velocity features of each joint of each skeleton are extracted, either regarding the joint or about the bone. The two types of velocities are concatenated at every moment and I represents their concatenation, with $I \in R^{2C\times V\times T}$.

3.3.2. Early fusion of Object knowledge & 3D Skeleton features

At this stage, S represents the position features of the skeleton data, I denotes the temporal features. For object knowledge, c is the position part, while d is the dynamic element. Two separated early fusions are made, S_c is the spatial early fusion by concatenating S and c, while I is concatenated with d as a temporal early fusion.

$$S_c = [S \quad c] \tag{6}$$

$$I_d = [I \quad d] \tag{7}$$

While $S_c \in R^{C1\times V\times T}$, $I_d \in R^{C2\times V\times T}$, with $C1 = N_p + 1$ and $C2 = N_d + 2$.

As the position representation S_c brings one information, while the temporal information I_d brings two (i.e. The class and the motion). To prevent any dimensional conflict, and since the motion of unmovable objects is always regarded as zero regardless of the scenario, another row is added to S_c, in turn, $s_c \in R^{C2\times V\times T}$.

After that, S_c and I_d are embedded separately into high dimensional space by two 1×1 convolution layers as follows (Equation (8) and Equation (9)):

$$\tilde{S_c} = Relu(w_2(Relu(w_1 S_c))) \tag{8}$$

$$\tilde{I_d} = Relu(w_4(Rely(w_3 I_d))) \tag{9}$$

Where $W_1, W_3 \in R^{C_3\times C_2}$ and $W_2, W_4 \in R^{C_3\times C_2}$, $Relu$ is the Relu activation function.
Thereafter, a summation joins Equation (8) and Equation (9) together:

$$Z = \tilde{S_c} + \tilde{I_d} \tag{10}$$

At this stage of the proceedings, Z in Equation (10) is on standby to be the input of the vision transformer network. To make the paper well-contained, we go through in detail in the following section the steps of the vision transformer to model this Z.

3.3.3. Training with Self-attention

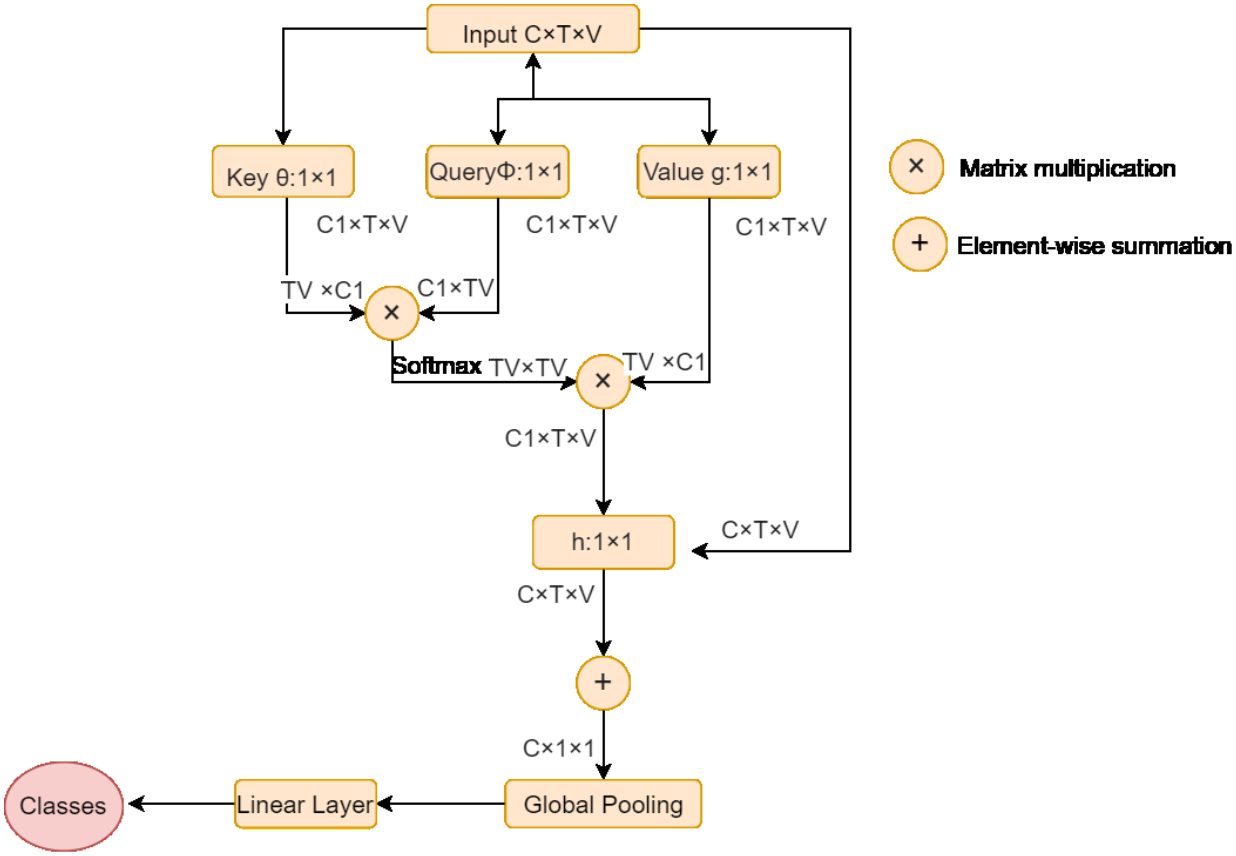

Figure 4. The Spatio-temporal self-attention block [13]

In our approach, the action recognition task is a supervised learning classification task. The objective is to acquire a robust representation of Z that leads to improved prediction accuracy for action classes. The incorporation of Object Knowledge into the skeleton representation resulted correspondingly, a bigger picture, which augmented the long-distance global dependency. On account of this, the spatial-temporal self-attention mechanism was the chosen model for this task to dynamically optimize the proposed structure (Figure 4), avoiding several times of parameters. Let x be the input of the network and y its corresponding output. The self-attention block is described as follows (Equation (11)):

$$y(x) = Relu(h(f(x).g(x))) + x \tag{11}$$

As $f(x)$ represents the similarity matrix (Equation (12)):

$$f(x) = softmax((w_\theta x)^T (w_\phi x)) \tag{12}$$

$g(x) = w_g x$, $h(x) = w_h x$, $\theta(x) = w_\theta x$, $\phi(x) = w_\phi x$ and w_h, w_θ, w_ϕ and w_g are learnable:1×1 convolution operation) and $+x$ represents a residual connection. The self-attention block produces feature maps in $R^{C \times T \times V}$ (Figure 4 represents the spatiotemporal self-attention block). Following the general architecture of the self-attention

architecture, a global pooling layer is applied, producing features in $R^{C\times1\times1}$. And finally, to generate the classes of actions, a linear layer is put in.

4. Experiments and Analysis

The proposed approach was extensively evaluated on different significant benchmark datasets, namely NTU RGB+D 60 [17], NTU RGB+D 120 [18] and PKU-MMD [19]. The analysis of variant models was performed to validate the contribution of each component, and a comparison with the state-of-the-art method was visualized to demonstrate the efficiency of the proposed solution.

4.1. Datasets and Protocols

4.1.1. NTU RGB+D 60 Dataset [17]

The NTU60 RGB+D dataset is an extensive and intricate compilation of 56,880 video clips that document 60 different human actions performed by 40 individuals, each with 25 joints represented in 3D coordinates and captured from 80 different perspectives using three Kinect cameras. The actions are categorized into three groups, namely daily, mutual, and health-related actions. The model evaluation was conducted following the experimental settings outlined in [36]. For the cross-subject (cs) evaluation, 20 of the 40 subjects were used for training, resulting in 40,320 videos for training and 16,560 videos for testing. The Cross View (cv) evaluation employed sequences captured by two cameras for training, while the remaining sequences captured by a third camera were reserved for testing. This resulted in 37,920 videos for training and 18,960 videos for testing.

4.1.2. NTU RGB+D 120 Dataset [18]

The NTU RGB+D 120 dataset is a continuation of the NTU RGB+D 60 dataset and consists of 114,480 skeleton sequences that represent 120 distinct action classes performed by 106 individuals. It is currently the most extensive dataset for recognizing human actions based on skeletons. Our evaluation methods are based on the standard protocols, including Cross-subject (cs) and Cross-view (cv). Under the CS protocol, we used videos from 53 human subjects for training and the remaining subjects for testing. For CV, we employed video clips captured by cameras with collection setup IDs for training and those captured by cameras with odd setup IDs for testing.

4.1.3. PKU-MMD Dataset [19]

The PKU-MMD dataset is a skeleton-based action recognition dataset featuring 51 actions. It contains 5,312,580 frames distributed across 3,000 minutes, with over 20,000 temporally localized actions. The dataset is categorized into two sections: 43 daily actions (e.g., drinking, waving hands, putting on glasses) and 8 interaction actions (e.g., hugging, shaking hands). The evaluation methods employed are Cross-subject (cs) and Cross-view (cv). For Cross-subject evaluation, 57 subjects are used for training and 9 for testing, with 944 and 132 long video samples in the respective sets. In Cross-view evaluation, the training and testing sets include 717 and 359 video samples, respectively.

4.1.4. Data Analysis

Table 2 and Table 3 illustrate the correlation between existing actions in NTU RGB+D and PKU-MMD respectively with the existing objects of the object 365 dataset [38]. For the NTU RGB+D dataset, 42.5% involve interactions with objects. However, when considering the availability of object detection datasets, the percentage of actions involving object interactions becomes 26.27%. For example, considering the action ”Put on jacket”, when looking at the Object 365 dataset, the ”Jacket is not included. This ratio indicates the portion of the data where the model incorporates object knowledge. For the PKU-MMD dataset, 30% of actions involve interactions with

Table 1. The various actions that involve interactions with objects are classified based on whether the objects are stationary or in motion, which determines the need for a study on the movement of objects of the NTU RGB+D 120 dataset

Number of the class	**Actions**	**Possible associated objects**	**Moving objects**
A1	Drink water	Cup	Yes
A2	Eat meal	Plate	Yes
A3	Brush teeth	Toothbrush	Yes
A11	Reading	Book	Yes
A12	Writing	Notepaper	No
A13	Tear up paper	Notepaper	Yes
A15	Shoot at basket	Basket	No
A16	Put on a shoe	Shoes	Yes
A17	Take off a shoe	Shoes	Yes
A18	Put on glasses	Glasses	Yes
A19	Take off glasses	Glasses	Yes
A20	Put on a hat/cap	Hat	Yes
A21	Take off a hat/cap	Hat	Yes
A28	Phone call	Telephone	Yes
A29	Play with phone/tablet	Tablet	Yes
A30	Type on a keyboard	Keyboard	No
A33	Check time	Watch	Yes
A61	Put on headphone	head Phone	Yes
A62	Take off headphone	headphone	Yes
A64	Bounce ball	Soccer	Yes
A65	Tennis bat swing	Tennis	Yes
A66	Juggle table tennis ball	Table tennis	No
A73	Staple book	Stapler	Yes
A76	Cutting paper	Scissors	Yes
A78	Open bottle	Bottle	Yes
A82	Fold paper	Notepaper	Yes
A83	Ball up paper	Notepaper	Yes
A87	Put on bag	Backpack	Yes
A88	Take off bag	Backpack	Yes
A90	Take object out of bag	Backpack	Yes
A94	Throw up cap/hat	Hat	Yes
A107	Wield knife	Knife	Yes
A113	Cheers and drink	Cup	Yes
A115	Take a photo	Camera	No

Table 2. The various actions that involve interactions with objects are classified based on whether the objects are stationary or in motion which determines the need for a study on the movement of objects of the PKU-MMD dataset

Actions	**Possible associated objects**	**Moving objects**
Taking a selfie	Telephone	Yes
Writing	Notepaper	No
Tear up paper	Notepaper	Yes
Make a phone call/answer phone	Telephone	Yes
Check time (from watch)	Watch	Yes
Playing with phone/tablet	Tablet	Yes
Reading	Book	No
Typing on a keyboard	Keyboard	No
Take off glasses	Glasses	Yes
Wear on glasses	Glasses	Yes
Take off a hat/cap	Hat	Yes
Put on a hat/cap	Hat	Yes
Drink water	Cup	Yes
Eat meal/snack	Cup	Yes
Brushing teeth	Toothbrush	Yes

objects. For the rest of the actions, the model relies on the skeleton features to predict the non-interacted actions or actions, or actions where objects don't exist in the objects detection dataset. This partitioning helps the model become less prone to overfitting.

4.2. Implementation Details

4.2.1. Data processing

For **the Object knowledge** task, Object365 [38] has been selected as the reference dataset for the object detection model due to its inclusion of approximately 365 objects, with around 50 such interactive objects. A pre-trained model is adapted for the object detection model without fine-tuning while 34 is exactly the number of considered objects. We select the highest detection scores of 25 objects, for both interactive movable objects and interactive non-movable objects ($N_d = N_p = V = 25$), if the number of detected objects in a sequence is less than 25, we increase its length to 25 by padding it with the existing objects in the sequence. For object-dynamic features, The Robust Scaling technique is used to normalize the motion of objects, ensuring that any outliers do not hurt the data. The resulting range of values after normalization is not constrained to a specific range but rather has a mean of zero and a standard deviation of one, making it suitable for the application. For **3D Skeleton data**, To ensure a fair comparison with the standard method, the paper's processing steps are applied [36]. Specifically, to make the skeleton data invariant to initial positions based on the first frame, we translate the data for each video clip. During the training phase, we down-sample the video sequence by dividing it into 20 segments ($k = 20$)., which is consistent across both datasets. During testing, 5 down-sampled sequences are randomly generated, following the same approach described in the referenced paper, and the mean score is calculated as the final action classification result. To improve the model's robustness, data augmentation is performed during training by randomly rotating the skeleton data by some degrees. The incorporation of object knowledge into the skeleton representation was carried

Table 3. Accuracies obtained on NTU RGB+d 60 dataset by different-level of the proposed method

	Components			Accuracy	
Baseline	Position	Dynamic	Object Knowledge	cv %	cs %
B0	×	×	×	96.1	90.5
B1	✓	×	×	96.4	90.6
B2	×	✓	×	96.8	90.9
B3	✓	✓	✓	**97.1**	**91.8**

out before the training phase, using the Numpy and Pandas libraries with thorough consideration taken regarding the size of matrices.

4.2.2. Experimental setup and parameter settings

The experiments were conducted using the PyTorch framework on a single NVIDIA GeForce GTX 1080 Ti. The Adam optimizer was employed, with an initial learning rate of 0.001, and the learning rate was decreased by a factor of 10 at the 60th, 90th, and 110th epoch for both datasets. The models were trained for 120 epochs using a label smoothing factor of 1.1 and cross-entropy loss for action classification. To prevent any gaps, it's important to ensure that the number of frames in the skeleton sequence matches the number of frames in the corresponding RGB video.

4.3. Ablation Study

We conducted a series of ablation studies on the NTU RGB+D 60 datasets to thoroughly evaluate the proposed method and examine the efficacy of its components. Specifically, our focus was on verifying the effectiveness of the object knowledge, position features, and dynamic features, which we assessed through these ablation studies. As shown in Table 3, we put four different methods, each with a different component of the proposed model. All baselines are defined as follows,

B0 A spatiotemporal network without the early fusion with objects features [13].

B1 spatial early fusion only: the baseline exclusively incorporates position features while disregarding the dynamic information.

B2 Temporal early fusion only: To examine the impact of dynamic features, we removed the object-position features from the object knowledge and kept only the object-dynamic features

B3 Spatial-Temporal Early fusion with object knowledge

The results indicate that the baseline (Efficient SAt) [13] improved in recognition performance when equipped with distinct components of the proposed methods (B1-B3). Thus, it can be inferred that incorporating object knowledge as a unified input along with a spatiotemporal one-stream network facilitates the model to acquire more sophisticated representations, leading to an improvement in performance. The findings reveal that, specifically, when tested individually, B2 exerts a greater impact on the model's performance than B1, implying that object motion is more pivotal than position features in skeleton-based action recognition. Nevertheless, the model attains its optimal performance when both are incorporated. An illustrative instance that highlights the significance of Dynamic Encoding is the model's ability to distinguish between the actions of writing on a notepaper and tearing it up, both of which involve interaction with the same object. However, the action of tearing up the paper entails a distinct motion of the paper in contrast to writing.

4.4. Comparison with the State-of-the-Art Approaches

By comparing the proposed method with several contemporary state-of-the-art approaches, it is shown how the fusion of object knowledge with skeleton representation can contribute to the effectiveness of human recognition tasks. Four different methods should be noted for evaluating the experiments of the current approach:

Table 4. Performance comparison with the state of the art on the NTU RGB+D 60 dataset

Type	Method	Year	cs(%)	cv (%)
M1	MST (joint) [40]	2021	89.0	95.1
	Double-head (joint) [41]	2021	90.3	96.1
	Efficient SAt (baseline) [13]	2022	90.5	96.1
M2	Deep LSTMZ+LSTM-AE [42]	2018	80.6	88.56
	RotClips+MTCNN [43]	2018	81.09	87.37
	AGC-LSTM [44]	2019	89.2	95.0
	AS-GCN [45]	2019	86.8	94.2
	CA-GCN [46]	2020	86.5	94.1
	MS-G3D [47]	2020	91.5	96.2
	Shift-GCN [48]	2020	90.7	96.5
	SGN [36]	2020	89.0	94.5
	MST (2s) [40]	2021	91.5	96.6
	Double-head(2s) [41]	2021	91.7	96.5
	Ta-CNN [49]	2022	90.7	95.1
	Angular-Encoding [50]	2022	91.6	96.3
M3	Multi-stream Interaction [15]	2022	91.5	96.5
M4	A joint learning of HOI and AR [16]	2022	90.0	95.7
(Ours)	**SAt-Object Integration**	-	**91.8**	**97.1**

Table 5. Performance comparison with the state of the art on the NTU RGB+D 120 dataset on top-1 accuracy. HOI: human object interaction and AR: Action recognition

Type	Method	Year	cs(%)	cv (%)
M1	MST (joint) [40]	2021	82.8	84.5
	Js DualHead-Net [41]	2021	84.6	85.9
	Efficient SAt [13] (baseline)	2022	85.7	86.8
M2	2s-AGCN [51]	2019	82.50	84.90
	Shift-GCN [48]	2020	85.9	87.6
	MS-G3D [47]	2020	86.9	88.4
	SGN [36]	2020	79.2	81.5
	AMCGC-LSTM [52]	2020	79.70	80.00
	2s ST-TR [53]	2021	85.10	87.10
	MST(2s) [40]	2021	87.5	88.8
	4sDualHead-Net [41]	2021	88.2	89.3
	4s STF-Net [42]	2022	85.10	87.10
	Ta-CNN [49]	2022	85.7	87.3
M3	Multi-stream Interaction [15]	2022	88.18	89.41
M4	A joint learning of HOI and AR [16]	2022	-	-
(Ours)	**SAt-Object Integration**	-	**88.61**	**90.6**

M1: Single-stream approaches without the object information integration.
M2: Multi-stream (spatiotemporal) methods dont́ include the object information.
M3: Multi-stream methods contain the object information.
M4: A joint learning technique of action recognition and human-object interaction.

Figure 5. Comparison results of different actions that interact with objects with (2s-AGCN) [51] and our method on the NTU RGB+D dataset. Table 2 contains a list of actions accompanied by their corresponding numbers for reference

While M1 involves processing an entire video using a single input stream, M2 utilizes separate spatial-temporal models, often convolutional networks, to extract features in multiple ways. Multi-stream architectures are generally more computationally demanding but can provide higher accuracy compared to single-stream architectures. However, a recent approach called the Efficient Self-Attention [13], has introduced a novel method for integrating different types of information, such as joint information, joint motion, bone information, and bone motion, into a single input for a self-attention mechanism. This approach has shown impressive results while requiring fewer computational resources. Building on this idea, the present approach fusionned the skeleton representation and the object knowledge information in early fusion before the training phase providing better. The method is based on a self-attention mechanism with five stacked self-attention blocks. The results have been obtained on three well-known datasets, namely NTU RGB+D 60 [17], NTU RGB+D 120 [18], and PKU-MMD [19].

4.4.1. NTU RGB+D 60

The baseline of this work is the Efficient Self-Attention [13], with reported performance in the paper of 90.5% (CS) and 96.1% (CV). Further integrating object knowledge into this framework improves to 91.8% (CS) and 97.1% (CV), surpassing the current state-of-the-art methods. Table 4 presents the reported comparison results. The limitations of the current state-of-the-art methods are typically observed in actions that entail interactions with objects due to the lack of complete information. However, our approach has effectively addressed this issue, as demonstrated by our experimental results. The proposed method outperforms M4 in accuracy with a margin of 1.8% and 1.4% in the cs metric and the cv metric respectively. Regarding the ”Type on a keyboard” action depicted in Figure 5, our approach demonstrated a precision rate of 90.1%, outperforming the 2s-AGCN baseline, which achieved an accuracy rate of 78%. These results illustrate that our method correctly leverages the benefits of object knowledge by integrating it into a unified network along with the skeleton representation.

Table 6. Performance comparison with the state of the art on the PKU-MMD dataset on top-1 accuracy. HOI: human object interaction and AR: Action recognition

Methods	Year	cs(%)	cv (%)
STA-LSTM [54]	2017	61.20	63.30
JCRRNN [55]	2016	64.60	66.90
Skeleton boxes [56]	2017	82.50	84.90
Li et al. [57]	2017	86.80	94.20
HCN [58]	2018	85.90	87.60
TSMF [59]	2021	95.8	97.8
MMNet [60]	2022	97.4	98.6
(Ours) SAt-Object Integration	-	**98.1**	**98.9**

4.4.2. NTU RGB+D 120

Table 5 contains the numerical results of the proposed method compared to the current state of the art on the NTU RGB+D 120 dataset. The proposed method outperforms the state-of-the-art. M3 employs three streams to investigate the movements of human skeletons, objects, and their interactions. Of note, during the training phase, the three streams are trained independently and subsequently integrated. However, the proposed method was able to achieve superior performance based on a single training phase, with a slight margin of advantage.

4.4.3. PKU-MMD

Despite the limited utilization of this dataset in previous research, our model showcased remarkable performance on it. As depicted in Table 6, we compared the accuracy of our model with that of other established models, revealing our model's superiority. Specifically, it achieved an impressive accuracy of 98.1% in CS and an even higher 98.9% in CV, demonstrating its strong capabilities and potential for further applications.

4.4.4. Per-class Accuracy Comparison

Figure 5 compares the per-class accuracy of our approach with the 2S-AGCN method, specifically for categories that require object interaction, listed on the x-axis concerning the NTU RGB+D 120 dataset. Notably, The proposed approach has significantly improved the accuracy of each mentioned class in comparison with 2S-AGCN. We have chosen to make the comparison per class with 2S-AGCN, as the results of this baseline are available on both NTU RGB+D 60 and NTU RGB+D 120 datasets. This comparison provides a detailed validation of the core principle of the paper by showcasing the impact of object integration. The proposed method primarily focuses on improving the accuracy of actions related to objects, while the accuracy of the remaining actions relies on the skeleton representation. It is notable that improving the accuracy of a composite of actions while keeping the rest the same leads to enhanced overall accuracy, as is intuitively evident.

4.4.5. Complexity Discussion

The efficiency of the model's baseline is emphasized by its comparably low number of parameters in comparison to other models. Notably, 0.98M parameters. Given that, the proposed work is founded on this model, this lightweight nature has been passed on to our model. **Since**: The proposed method incorporates five self-attention layers which leverage its capacity for global feature extraction. This enables the model to achieve good performance while using fewer stacked layers, unlike GCN models. The number of frames taken is 20 and the number of joints stays the same at 25 (25*20). keeping the advantage of the low-resolution image mentioned in the baseline. Most embedding operations are linear and achieved by 1×1 convolutions, and **Despite that**: The channel input size is changed to include the collaborative embedding of object knowledge with the skeleton representation data, the

impact of this modification on the model's lightweight nature is minimal. Since the fusion is before the training phase, the model incorporates RGB data as a pre-trained object detection model. This integration does not increase complexity.

5. Conclusion

The integration of the existing objects and their dynamic behavior as supplementary indicators for skeletal characteristics leads to favorable outcomes in terms of accuracy and complexity. This highlights the importance of examining changes in the surrounding environment complying with the human features for the action recognition task. Our approach to integrating the environment into skeleton characteristics is novel. We formulated object knowledge data obtained from a pre-trained object detection model to include the position and dynamic features of objects. Additionally, we utilized an embedding technique to fuse object knowledge and skeletal features into a single-stream network. The resulting model was then fitted into a self-attention network model to facilitate automatic learning of the features of this new representation. Experiments conducted on popular Skeleton datasets demonstrated the superiority of this method over the current state-of-the-art.

6. Acknowledgement

This work was supported by the Euro-Mediteranean University of Fez.

REFERENCES

1. Ramchandra Rimal. Identifying the neurocognitive difference between two groups using supervised learning. *Statistics, Optimization & Information Computing*, 12(1):15–33, 2024.
2. Youssef Ben Youssef, Mohamed Merrouchi, Elhassane Abdelmounim, and Taoufiq Gadi. Classification of aircraft in remote sensing images based on deep convolutional neural networks. *Statistics, Optimization & Information Computing*, 10(1):4–11, 2022.
3. Maxim Zozyuk, Dmitri Koroliouk, Pavel Krysenko, Alexei Yurikov, and Yuriy Yakymenko. Prediction of characteristics using a convolutional neural network based on experimental data on the structure and composition of metamaterials. *Statistics, Optimization & Information Computing*, 2023.
4. Yu Kong and Yun Fu. Human action recognition and prediction: A survey.
5. Li Chen, Nan Ma, Patrick Wang, Jiahong Li, Pengfei Wang, Guilin Pang, and Xiaojun Shi. Survey of pedestrian action recognition techniques for autonomous driving. 25(4):458–470.
6. Sharath Chandra Akkaladevi and Christoph Heindl. Action recognition for human robot interaction in industrial applications. In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pages 94–99. IEEE.
7. Chhavi Dhiman and Dinesh Kumar Vishwakarma. A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence*, 77:21–45, 2019.
8. Andrea Pennisi, Domenico D Bloisi, and Luca Iocchi. Online real-time crowd behavior detection in video sequences. *Computer Vision and Image Understanding*, 144:166–176, 2016.
9. Ziqi Li and Dongsheng Li. Action recognition of construction workers under occlusion. *Journal of Building Engineering*, 45:103352, 2022.
10. Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE international conference on multimedia & expo workshops (ICMEW)*, pages 597–600. IEEE, 2017.
11. Xinghao Jiang, Ke Xu, and Tanfeng Sun. Action recognition scheme based on skeleton representation with ds-lstm network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2129–2140, 2019.
12. Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022.
13. Xiaofei Qin, Rui Cai, Jiabin Yu, Changxiang He, and Xuedian Zhang. An efficient self-attention network for skeleton-based action recognition. *Scientific Reports*, 12(1):4111, 2022.
14. Rongjie Xia, Yanshan Li, and Wenhan Luo. Laga-net: Local-and-global attention network for skeleton based action recognition. *IEEE Transactions on Multimedia*, 24:2648–2661, 2021.
15. Haoran Wang, Baosheng Yu, Jiaqi Li, Linlin Zhang, and Dongyue Chen. Multi-stream interaction networks for human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):3050–3060, 2021.
16. Liang Xu, Cuiling Lan, Wenjun Zeng, and Cewu Lu. Skeleton-based mutually assisted interacted object localization and human action recognition. *IEEE Transactions on Multimedia*, 2022.
17. Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

18. Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019.
19. Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017.
20. Raviteja Vemulapalli, 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*, pages 588–595, 2014.
21. Hanbo Wu, Xin Ma, and Yibin Li. Spatiotemporal multimodal learning with 3d cnns for video action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1250–1261, 2021.
22. Avinandan Banerjee, Pawan Kumar Singh, and Ram Sarkar. Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition. *IEEE transactions on circuits and systems for video technology*, 31(6):2206–2216, 2020.
23. Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1915–1925, 2020.
24. Shuangyan Miao, Yonghong Hou, Zhimin Gao, Mingliang Xu, and Wanqing Li. A central difference graph convolutional operator for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4893–4899, 2021.
25. Zhengcen Li, Yueran Li, Linlin Tang, Tong Zhang, and Jingyong Su. Two-person graph convolutional network for skeleton-based human interaction recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
26. Cong Wu, Xiao-Jun Wu, and Josef Kittler. Graph2net: Perceptually-enriched graph learning for skeleton-based action recognition. *IEEE transactions on circuits and systems for video technology*, 32(4):2120–2132, 2021.
27. Zengxi Huang, Yusong Qin, Xiaobing Lin, Tianlin Liu, Zhenhua Feng, and Yiguang Liu. Motion-driven spatial and temporal adaptive high-resolution graph convolutional networks for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1868–1883, 2022.
28. Sangwoo Cho, Muhammad Maqbool, Fei Liu, and Hassan Foroosh. Self-attention network for skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 635–644, 2020.
29. Yanbo Fan, Shuchen Weng, Yong Zhang, Boxin Shi, and Yi Zhang. Context-aware cross-attention for skeleton-based human action recognition. *IEEE Access*, 8:15280–15290, 2020.
30. Mrugendrasinh Rahevar, Amit Ganatra, Tanzila Saba, Amjad Rehman, and Saeed Ali Bahaj. Spatial–temporal dynamic graph attention network for skeleton-based action recognition. *IEEE Access*, 11:21546–21553, 2023.
31. Jianbo Liu, Ying Wang, Shiming Xiang, and Chunhong Pan. Han: An efficient hierarchical self-attention network for skeleton-based gesture recognition. *arXiv preprint arXiv:2106.13391*, 2021.
32. Kai Hu, Junlan Jin, Chaowen Shen, Min Xia, and Liguo Weng. Attentional weighting strategy-based dynamic gcn for skeleton-based action recognition. *Multimedia Systems*, pages 1–14, 2023.
33. Jiaxu Zhang, Gaoxiang Ye, Zhigang Tu, Yongtao Qin, Qianqing Qin, Jinlu Zhang, and Jun Liu. A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition. *CAAI Transactions on Intelligence Technology*, 7(1):46–55, 2022.
34. Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.
35. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
36. Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1112–1121, 2020.
37. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
38. Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.
39. Tsung-Yi Lin, 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*, pages 740–755. Springer, 2014.
40. Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1113–1122, 2021.
41. Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Yu Guan, Xuming He, and Errui Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4334–4342, 2021.
42. Juanhui Tu, Hong Liu, Fanyang Meng, Mengyuan Liu, and Runwei Ding. Spatial-temporal data augmentation based on lstm autoencoder network for skeleton-based human action recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3478–3482. IEEE, 2018.
43. Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning clip representations for skeleton-based 3d action recognition. *IEEE Transactions on Image Processing*, 27(6):2842–2855, 2018.
44. Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1227–1236, 2019.
45. Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019.

46. Xikun Zhang, Chang Xu, and Dacheng Tao. Context aware graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14333–14342, 2020.
47. Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.
48. Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 183–192, 2020.
49. Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2866–2874, 2022.
50. Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, RI McKay, Saeed Anwar, and Tom Gedeon. Fusing higher-order features in graph neural networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
51. Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.
52. Shihao Xu, Haocong Rao, Hong Peng, Xin Jiang, Yi Guo, Xiping Hu, and Bin Hu. Attention-based multilevel co-occurrence graph convolutional lstm for 3-d action recognition. *IEEE Internet of Things Journal*, 8(21):15990–16001, 2020.
53. Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pages 694–701. Springer, 2021.
54. Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
55. Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. Online human action detection using joint classification-regression recurrent neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 203–220. Springer, 2016.
56. Bo Li, Huahui Chen, Yucheng Chen, Yuchao Dai, and Mingyi He. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 613–616. IEEE, 2017.
57. Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 585–590. IEEE, 2017.
58. Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018.
59. XB Bruce, Yan Liu, and Keith CC Chan. Multimodal fusion via teacher-student network for indoor action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3199–3207, 2021.
60. XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022.