Multi Modal Aware Transformer Network for Effective Daily Life Human Action Recognition

Hend Basly[1(✉)], Mohamed Amine Zayene[1,2], and Fatma Ezahra Sayadi[1]

[1] NOCCS-Lab.: Networked Objects Control and Communication Systems Laboratory, University of Sousse, National Engineering School of Sousse (ENISO), Tunisia National Engineering School of Sousse, University of Sousse, BP 264 Erriadh, Sousse 4023, Sousse, Tunisia
{basly.hend,Fatmazahra.sayadi}@eniso.u-sousse.tn
[2] Faculty of Sciences, University of Monastir, Monastir, Tunisia
mohamedamine.zayene@fsm.rnu.tn

Abstract. Multimodal approaches for human action recognition leverage the complementary nature of different data modalities to improve system performance in healthcare, sports, smart homes, and surveillance. However, effectively fusing features from different modalities is challenging. Generally, action recognition involves extracting spatial and temporal information from video footage. To this end, 3D poses are fed into an STGCN to effectively capture spatial relationships between body parts and the temporal dependencies between consecutive frames. Self-attention was applied in the spatial dimension to focus on joint relationships dynamically in the same frame. External attention was used in the temporal dimension to capture relevant temporal dynamics between different joints over time. A 3D CNN was applied to extract spatial and temporal information from RGB data. Both were combined using the Intermodal Attention-based Fusion (IAF) component, which makes it possible to take into account the context information modalities by enabling the model to concentrate on the highly informative aspects of each modality during fusion. By learning attention weights, the model can selectively weight the contributions of different modalities based on their relevance to the action being recognized. Tested on the NTU-RGB+D dataset, the architecture outperformed existing approaches, showing promising effectiveness.

Keywords: Self-Attention Mechanism · External-Attention Mechanism · Daily Life Action Recognition

1 Introduction

In this paper, we delve into the domain of human activity recognition, a broad field that deals with videos sourced from diverse origins, including films, recordings,

Supported by organization x.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024, LNCS 14813, pp. 165–179, 2024.
https://doi.org/10.1007/978-3-031-64605-8_12

surveillance cameras, and web content. Recognizing actions in videos, particularly in the context of Daily Life Actions (DLAs), poses a set of challenges. DLA videos require a different approach as contextual information becomes less significant, making the task more complex. DLA videos exhibit distinct characteristics that increase their complexity. These challenges encompass the need to develop camera view-invariant representations, model temporal information effectively, handle variations among actions that exhibit resemblances (analogous actions), recognize actions presenting fine details (fine granularity actions), and manage substantial variability intraclass within one same activity category. Covering a diverse set of actions, particularly focusing on the context of DLA, demands more robust action recognition algorithms. One potential solution to these challenges involves employing 3D convolutional neural networks (3D ConvNets) for video analysis. However, it is crucial to question the effectiveness of these networks for DLA recognition. Existing networks like I3D [1], ResNet [2] and ResNeXt [3] were primarily predestined to handle generic video structures, not the specific spatiotemporal cubes encountered in DLAs. They lack the capacity to capture long-term temporal relationships, a vital aspect of tackling DLA challenges. While some recent works have highlighted the efficiency of 3D convolutional operations in modeling temporal information, the intricate nature of DLA recognition requires additional skills beyond fixed convolutional operations. To meet these challenges comprehensively, we introduce leveraging multiple modalities to harness the complementary nature of their features, thus obtaining the most discriminative information for each action. Utilizing 3D ConvNets allows us to capture short-term motion and appearance information without relying on optical flow, which is less relevant in DLAs that primarily involve subtle motions. However, combining 3D skeleton data with different stream networks can lead to overfitting due to the high parameter count in multimodal networks. To address these issues, we turn to Graph Convolutional Networks (GCNs), which effectively represent 3D skeleton data. Spatio-Temporal Graph-Convolutional-Networks (STGCNs) gather data pertaining to both spatial and temporal aspects by examining bone connections in 3D skeleton data. Yet, they have limitations related to their graph structure. In the past, Recurrent Neural Networks (RNNs) were commonly used for modeling temporal dependencies. However, RNNs have a limited window to reference from, which becomes problematic with longer videos. To address these limitations, we turn to transformers, basically developed for Natural Language Processing (NLP). Transformers, making use of their self attention mechanism, can effectively model long-term temporal dependencies and are well-suited for the hierarchical and sequential nature of skeletal sequences. While self-attention mechanisms are effective for capturing local relationships within data samples, they may fall short in fully capturing the inter-joint dynamics over a sequence of skeletal data. Therefore, we introduce the concept of external attention to address this limitation. Our proposed framework combines the strengths of these approaches by innovative approach combining 3D poses in a self attention mechanism unit in the spatial dimension and an external attention mechanism unit in the temporal dimension to influence RGB data through Intermodal Attention-based Fusion (IAF) module. Thus, we employ two separate spatial an temporal streams. The spatial stream focuses

on the local relationships among skeleton joints in the context of every individual frame. Meanwhile, the temporal stream detects global temporal connections among joints across time, depicting action dynamics across the video sequence. We then enhance our proposal by fusing the significant spatial-temporal data of the 3D pose data with spatiotemporal visual information in the RGB data using the Intermodal Fusion (IAF) component. This fusion process extracts crucial spatial and temporal information from the 3D pose graph representation and the 3D visual feature representation from RGB clips, ultimately contributing to more effective DLA recognition in real-world scenarios. The remainder of this paper is structured as follows; in Sect. 2, we present the related work. In Sect. 3, we present our suggested Action Recognition methodology. In Sect. 4, we provide our experimental analysis conducted on the publicly accessible NTU-60 dataset and finally, we conclude in Sect. 5.

2 Related Work

Human action recognition has received greater attention in the field of computer vision and pattern recognition owing to its large variety of applications such as video monitoring, human-computer interaction and activity analysis.

This section reviews the existing literature and research contributions in this field, aiming to provide an overview of the latest techniques and advancements. We begin by discussing early works laid the foundation for progress in human action recognition, which has seen significant advancements in recent years, largely driven by deep learning. Convolutional neural networks (CNNs) have been a key tool, with notable studies introducing two-stream CNN architectures for capturing appearance and motion information [4,5]. However, these models are limited to 2D inputs and overlook motion patterns across multiple frames. For DLA recognition, Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have been used to model temporal relationships [6]. Combining CNNs and LSTMs has gained interest, effectively addressing spatial and temporal feature extraction [6]. However, they often rely on strong human motion videos, impacting their performance on DLA videos [7,8]. 3D CNNs have emerged as effective models for learning representations from volumetric data, extracting both appearance and motion information across contiguous frames [1]. However, they require pretraining on large datasets and only cover a small portion of video sequences, limiting their ability to capture long-range temporal dynamics. Graph Neural Networks (GNNs), specifically Graph Convolutional Networks (GCNs), have been employed to recognize actions from skeleton data, effectively capturing spatial and temporal dependencies between joints [9]. Several architectures, like DPRL [10] and DGNN [11], have leveraged GCNs for Human Activity Recognition. The ST-GCN model extends GCN to represent skeletal data and transform joints and bones into nodes and arcs, respectively [12]. While Both approaches relying on RGB data and those relying on 3D pose information have their merits, the latter excels in actions characterized by Notable human motion and resemblance in appearances.

The representation of temporal information has evolved, with RGB methods using spatio-temporal convolutions and 3D pose methods adopting spatio-temporal graph-based operations. Attention mechanisms have gained attention recently, improving action recognition accuracy by selectively focusing on important spatial or temporal regions within sequences [13–17]. These mechanisms have been employed in various ways, such as multilayer LSTM networks with spatio-temporal attention [13], attention clusters with shift operations [14], and global and local knowledge-aware attention networks [18]. In the following sections, we present our proposed methodology for human action recognition, taking inspiration from the advancements discussed in this "Related Work" section.

3 Proposed Architecture

In this section, we describe our proposed deep network DLA recognition. For a better video's analysis and understanding, we opted for the video action representation from multimodal data to reap the benefits of RGB data and skeleton, hence, acquire more distinct representations for intricate scenarios in Daily Life Actions (DLAs). To this end, we propose an innovative methodology based on Spatial Self Attention module-based transformer and Temporal External Attention module in conjunction with Highly effective 3D convolutional network. Extracted features are fused by an attention-based inter-modal fusion module, which enables the model to focus on the most informative parts of each modality during the fusion task. By learning attention weights, the model can selectively weight the contributions of different modalities based on their relevance to the activity being recognized. This allows for a more fine-grained fusion that takes into account the semantic relationships between modalities. The inputs of the proposed model are the RGB data frames of each video for the 3D CNN network and their corresponding 3D skeleton data modality for the spatial and temporal streams. We have applied data augmentation operations as center crops, random crops and horizontal and vertical random flips. A 4D feature map was extracted from the layer preceding the Global Average Pooling (GAP) layer of the RGB stream (I3D network) to represent spatial-temporal features. Both spatial and temporal streams are based on STGCN network for modeling spatio-temporal data. Its architecture is composed of nine layers, see [20] to have clear sight. In fact, STGCN combines two main components which are; Graph Convolutional Network (GCN) followed by Temporal Convolutional Network (TCN). GCN layers are typically used to capture the spatial relationships among different locations in the data and TCN layers are used to model the temporal aspects of the data, such as how the data evolves or changes over time. STGCN effectively captures both the spatial and temporal dependencies in spatio-temporal data. This makes it a powerful architecture for DLA recognition task.

In our architecture, from the 4^{th} layer, we apply Spatial Self Attention Unit (SAU) instead of GCN for the spatial stream and we apply Temporal External Attention Unit (TEU) instead of TCN in the temporal stream. The results of spatial and temporal streams are respectively the spatial and temporal characteristics of the skeleton modality, taking into account spatial and temporal

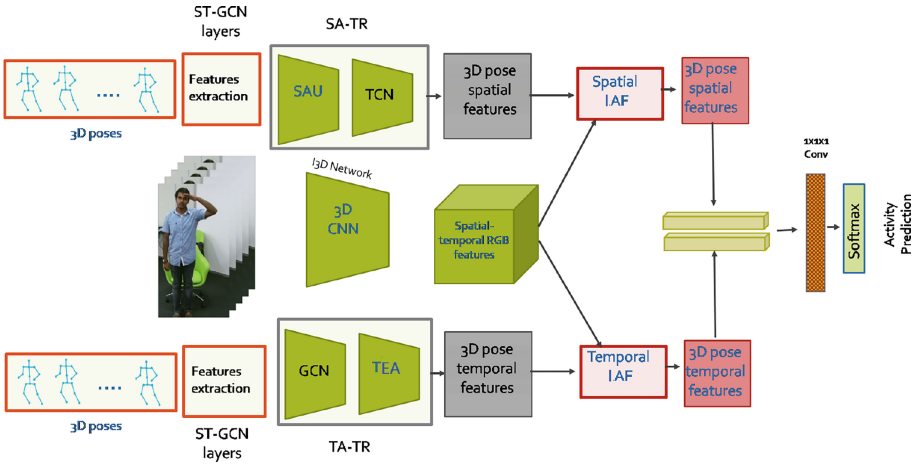


Fig. 1. End-to-End Spatial-Temporal Attention Network for Action Recognition

dependencies that have greater contribution to the realization of the performed action. The spatial extracted features from 3D poses and the spatio-temporal extracted features from RGB clips (see Fig. 1) are fused through the Intermodal Attention Fusion (IAF) module and produces new spatial features. This module allows the features extracted from skeleton data to re-calibrate features extracted from RGB data allowing the model to take into account contextual information to focus on the most informative parts of each modality during fusion. Similarly, The temporal attention features of the skeleton modality and the spatio-temporal features of the RGB modality (see Fig. 1) are fused through the Intermodal Attention Fusion (IAF) module to produce new temporal features. New spatial and temporal resulting features are concatenated and passed by softmax activation function to predict actions.

3.1 Spatial Self Attention Unit (SAU)

The SAU unit is typically employed to spotlight local correlations between different spatial body joints within each frame of a video or spatiotemporal sequence. In SAU, each frame is treated as an independent entity, and the goal is to extract low-level features that encode the interconnections and dependencies between various body joints. This is achieved through the use of self attention mechanism a per-frame basis, operating between adjacent skeleton joints, enabling the model to assess the importance of relationships between each pair of skeleton joints. In fact, part of the transformer encoder, the SAU unit converts a fully connected body skeleton node graph into a weighted graph, with weights reflecting the importance of intra-frame joint connections. These weights are then employed to evaluate each node's impact on the performed action. Essentially, SAU is the outcome of multiple parallel attention heads, each utilizing self attention to capture correlations among input sequence embeddings.

Given a sequence of n nodes of the skeleton, denoted as $ns_i^f \in R^{C_{inp}}$ for $i = (1, \ldots, E_v)$ embeddings, where C_{inp} represents the channel input for each frame $f \in F = 1, ..., N$, and N represents the overall count of frames. For every graph's node ns_i^f, we calculate $q_i^f \in R^{dq}$, $q_i^f \in R^{dq}$ and $q_i^f \in R^{dq}$, representing the query matrix, the key matrix and the value matrix respectively, through linear transformations of the joint data. Subsequently, for every pair of nodes (nd_i^f, nd_j^f), we compute a score $\alpha_{ij} \in R$ which quantifies the importance of associations between the two body joints, expressed through $\alpha_{ij}^f = q_i^f k_j^{f^F}$ for all $f \in F$. The resulting score is then applied to to assign weights to each joint, then the score is multiplied by the value of the joint v_j^f. In order to assess the overarching node's importance ns_i within a sequence, as depicted in Fig. 2, we calculate the weighted node Z_i^t using the softmax activation function, expressed as follows:

$$Z_i^f = \sum_j softmax\left(\frac{\alpha_{ij}^f}{\sqrt{d_k}}\right) V_j^f \tag{1}$$

The spatial stream (SA-TR) employs the principle of transformer network on 3D pose data using SAU. The self attention mechanism is applied instead the sub-module GCN through SAU to extract spatial relevant features.

3.2 Temporal External Attention (TEA) Unit

The TEA unit focuses on studying how each body joint behaves independently across multiple frames in the sequence. It treats each joint as a distinct entity and explores how these joints evolve over time, considering the temporal dimension. Unlike traditional self-attention mechanisms, TEA introduces an external memory unit that stores information relevant to these body joints [19]. TEA allows the model to dynamically adapt and learn temporal dependencies and interactions between different joints across various time steps. For example, it can uncover how specific body joints in one frame correlate with the same joints in subsequent frames, enabling the capture of complex and discriminative features. This adaptability is particularly valuable when analyzing human body movements, as it can reveal intricate relationships and dynamics that may be challenging to capture with conventional convolutional approaches using fixed kernel sizes. The following equation is used to calculate the temporal attention scores (alpha) based on the interaction between temporal features ($\mathbf{F}_i$) at time i and the information stored in the external memory M_k at position j.

$$\alpha_{i,j} = \exp\left(\frac{\mathbf{F}_i \cdot \mathbf{M}_j^T}{\sum_k \exp(\mathbf{F}_i \cdot \mathbf{M}_k^T)}\right) \tag{2}$$

The softmax function is employed to normalize these scores, ensuring that they sum to 1.

$$\alpha'_{i,j} = \frac{\exp(\alpha_{i,j})}{\sum_k \exp(\alpha_{k,j})} \tag{3}$$

Additional normalization is performed to provide a meaningful interpretation of the scores.

$$\alpha_{i,j} = \frac{\alpha'_{i,j}}{\sum_k \alpha'_{k,j}} \tag{4}$$

Temporal features are then weighted according to the attention scores $\alpha_{i,j}$ and the M_v matrix, emphasizing the most significant frames of the temporal sequence for activity classification. Just like the (SA-TR) stream, the (TA-TR) can be understood as the conventional graph convolution sub-module (GCN) [12] followed by the TEA unit instead of TCN. The spatial and temporal attention units (SAU and TEA) are integrated into the encoder module of the transformer network through the use of the multi-headed attention sub-module.

3.3 Intermodal Attention-Based Feature Fusion Module (IAF)

The Intermodal Attention-based Fusion (IAF) module depicted in Fig. 2 is designed to fuse information from the 3D poses' features extracted from the spatial and the temporal models into RGB's features extracted from the convolutional model. To take into account multimodal interactions between RGB and Skeleton data, we employ bilinear pooling operation. This specific application is particularly useful for fine-grained action classification.

$Feat_{rgb}$ and $Feat_{ske}$ are the characteristics Derived from the RGB stream and skeleton stream (in spatial domain or temporal domain), respectively. We apply the bilinear pooling operation, the resulting vector of feature combination is:

$$F_C = V_{AP}(Feat_{rgb} \otimes Feat_{ske}) \text{ where } F_C \in R^{M\times 1} \text{ and } M = Ch \times Ch \tag{5}$$

where Ch indicates the feature channels' number, we use the average pooling vector denoted as V_{AP} to transform the input features into vectors, ensuring that they all share identical dimensions. The symbol $\otimes$ signifies the outer-product operation performed on resultant matrices. Subsequently, we apply l2 normalization to regularize the F_C vector, and this is achieved as follows:

$$F_C = \text{sign}(F_C) \cdot \sqrt{|F_C|} \text{ where } F_C \in R^{M\times 1} \tag{6}$$

$$F_C = \frac{F_C}{\|F_C\|_2} \tag{7}$$

Following this, F_C is propagated via a fully-connected layer and subsequently via the ReLU layer to introduce nonlinearity, enabling our model to learn additional features. This process results in the generation of a new combined feature, denoted as $F_C^{'} \in R^{Ch\times 1}$, represented as:

$$F_C^{'} = \text{ReLU}(W \cdot F_C) \text{ with } W \text{ representing the weight matrix.} \tag{8}$$

The fusion process continues as follows; The resultant fusion feature undergoes a 1D convolution layer, followed by the application of the sigmoid activation function to generate new weights. This procedure enhances the importance of specific features, yielding the final fusion features for both the RGB and skeleton modalities, as described by the following equations:

$$Y_T = Feat_{rgb} * \sigma(\text{conv1D}(F_C^{'})) \tag{9}$$

The "*" operator represents the Hadamard product, while σ denotes the sigmoid activation function. The outcome is a representation of the scaled weights, reflecting features' importance. The weight values are then used to multiply the initial feature map, resulting in new weights denoted as Y_T. These updated weights signify the reconfiguration of the original features Gathering the underlying semantic insights pertaining to identical human actions expressed across diverse data modalities. The IAF is employed to explore possible semantic relationships among different modalities of data.

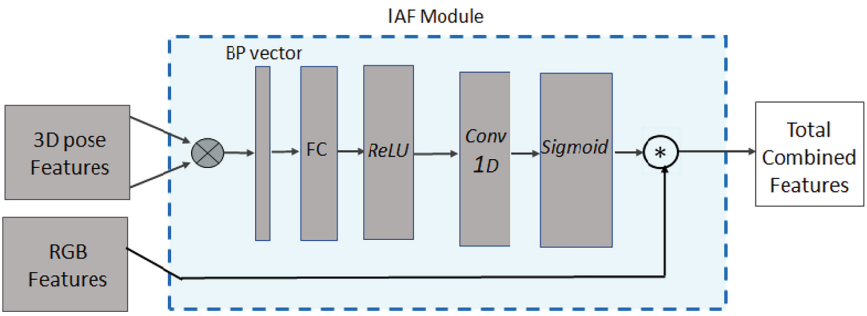


Fig. 2. IAF Module.

4 Model Evaluation Procedure

We conduct a series of experiments within our suggested framework to assess the effectiveness of the proposed architectural modules, particularly their impact on recognition performance. We evaluate our model's performance and compare it with several recent methods using the dataset [22], which is known for its complexity and extensive scope in the realm of DLA recognition. It's a large-scale resource capturing RGB, depth, skeleton, and infrared data from diverse action classes executed by 40 participants and recorded from 80 unique perspectives. This dataset complies with two suggested evaluation procedures, Cross-View and Cross-Subject, making it a robust platform for our evaluation.

4.1 Implementation Details

In our implementation, we carefully follow a series of steps to train our multi-modal action recognition network. Here is a summary of the key implementation details, including the algorithm steps:

1. Initialization and Pretraining: We utilize the PyTorch deep learning framework for our experiments. The I3D network is pretrained on the ImageNet dataset, and its initial weights are set accordingly.
2. Data Augmentation and I3D Training: Data augmentation techniques are applied, and the I3D network is trained on consecutive human body segments extracted from RGB video frames. This procedure entails the random selection of an initial image from the first half of the sequence and then including subsequent frames at intervals of 2.
3. Fine-Tuning I3D Backbone: The network undergoes separate fine-tuning using the SGD optimizer on the NTU—60 benchmark dataset. Key hyperparameters include a learning-rate equal to 0.01, a weight-decay equal to 0.0001, and drop-out-rate of 0.5.
4. Feature Extraction: We extract RGB feature maps from the Mixed-5c-layer of our backbone, yielding features with dimensions of (1024 channels $\times$ 8 frames $\times$ 7 $\times$ 7 pixels).
5. Training Spatial and Temporal attention Models: Our Spatial and Temporal skeleton branches are trained using sequences of 3D pose sampled from video clips. The sequence length (L) is 300, the training involves 120 epochs with a batch size of 32. We use the SGD optimizer with an initial learning rate of 0.1, applying learning rate reduction by a factor of 10 at specified epochs.
6. Training Complete Network for Classification: For the classification task, we incorporate a Global Average Pooling (GAP)layer, apply a drop-out rate of 0.5, and include a Soft-Max layer towards the final stages of the network. Hyperparameters γ_1 and γ_2 are set to 10^{-5}. The complete training process spans thirty epochs and utilizes the SGD optimizer method. It commences with an initial-learning rate of 10^{-3}, employing learning-rate-decay every ten epochs. The batch size remains constant at 32 throughout the training.

Our implementation runs on 4 Titan RTX graphics processing units (GPU).

4.2 Comparison of Proposed Methodology and State of the Art Approaches on NTU— RGB+D Dataset

We evaluate our STA-IAF model against the most current and advanced techniques using the NTU dataset, presenting precision measurements for each of the NTU-Cross-Sub and the NTU-Cross-View evaluation protocols, as depicted in Table 1. Remarkably, STA-IAF surpasses the performance of existing approaches. In comparison to works like [20,21], which rely solely on 3D pose data as input, STA-IAF achieves a classification accuracy that closely matches theirs for the NTU-Cross-View protocol, alongside marginal differences of; 1% compared to [20] and 0.9% compared to [21]. This closeness in accuracy can be attributed

Table 1. Accuracy Comparison on NTU— RGB+D Dataset using NTU-Cross-Sub and NTU-Cross-View Evaluation Protocols. '+' indicates that the specified modality type was exclusively employed during the training stage. '*' designates that the checked work has employed the depth modality type.

Method	Modality		Attention		Protocol	
	RGB	3D pose	with	without	NTU-Cross-Sub	NTU-Cross-View
S, Yan et al. (2018) [12]	–	x	–	x	81.5%	88.3%
F, Baradel et al. (2017) [23]	x	x	x	–	82.5%	88.6%
F, Baradel et al. (2018) [24]	x	x	x	–	84.8%	90.6%
G, Liu et al. (2019) [25]	x	x	x	–	85.4%	91.6%
F, Baradel et al. (2018) [26]	x	+	x	–	86.6%	93.2%
L, Shi et al. (2018) [27]	–	x	x	–	88.5%	95.1%
F, Shi et al. (2021) [28]	–	x	x	–	83.4%	89.0%
Y, Tang et al.(2018) [10]	–	x	x	–	83.5%	89.8%
R.M.R, Guddeti et al^{+} (2024) [37]	–	x	x	–	83.8%	89.75%
C, Li et al. (2018) [29]	–	x	–	x	86.5%	91.9%
M, Li et al. (2019) [30]	–	x	x	–	86.8%	94.2%
S, Cho et al. (2020) [31]	–	x	x	–	87.2%	92.7%
Y, Sun et al. (2021) [32]	–	x	x	–	88.4%	93.2%
C, Plizzari et al. (2020) [20]	–	x	x	–	89.9%	96.1%
Z, Zhang et al. (2020) [33]	–	x	x	–	89.3%	95.8%
L, Shi et al. (2019) [11]	–	x	–	x	89.9%	96.1%
HRV, Joze et al. (2020) [22]	x	x	x	–	91.9%	95.3%
S, Cho et al. (2020) [21]	–	x	–	x	91.5%	96.2%
T, Ahmad et al. (2023) [36]	–	x	x	–	91.5%	96.8%
M, Liu et al. (2018) [34]	x	x	x	–	91.7%	95.2%
S, Das et al. (2019) [35]	x	x	x	–	92.2%	94.6%
STA-IAF (ours)	x	x	x	–	**94.8%**	**97.1%**

to their robustness in handling view variations and capturing motion patterns. However, it's worth noting that these pose-only approaches struggle with distinguishing actions that have very similar visual appearances, leading to lower precision in the NTU-Cross-Sub protocol.

Furthermore, in comparison to other methods employing 3D poses and attention mechanisms for fusion, such as 2 s-AGCN [27], STAR [10,20,28,30–32], our approach, which combines bilinear pooling and attention mechanism demonstrates superior performance, as showcased in Table 1. These results validate the potential and generalization capability of our proposed approach. Additionally, our STA-IAF proposed model is compared with previous approaches implementing 3D poses, RGB and attention mechanism such as [22–25,34,35], STA-IAF outperforms all these methods. The outstanding performance of STA-IAF highlights the efficacy of our suggested methodology, particularly in generating adaptable data representations for different views. By combining transformer concept into the STGCN and incorporating Intermodal fusion through the external attention module (IAF), our model excels in capturing long-term

relationships from 3D poses data and effectively merges them with visual appearance information from the RGB data. In conclusion, our STA-IAF model, with the integration of the spatial self attention module, the temporal external attention module and the intermodal fusion strategy, excels in extracting significant and discriminative spatial-temporal features essential for actions of daily life (DLA) recognition.

4.3 Ablation Study

In this part, we make an ablation study on the proposed STA-IAF to justify the effectiveness of this framework with NTU— RGB+D dataset.

Effect of Associating or Dissociating Spatial and Temporal Attention Mechanisms. We test the robustness of the model first by dissociating the attention modules, each in a single one, (spatial (SA-TR) or temporal (TA-TR)) and second by associating them. Table 2 displays the performance comparison of all the implemented configurations. The first line shows results for the I3D base network without attention mechanism (without-TR); the second one shows the results of the combination of the I3D base network and the STGCN model through IAF model (STGCN-I3D-IAF). The third row shows the results of applying the spatial self attention (SA-TR) and then the temporal external attention (TA-TR) inside the STGCN network and combining the result with I3D by the IAF fusion module. The forth is intended only for the spatial self attention (SA-TR-I3D-IAF) which consists of combining the spatial self attention module by I3D and fuse them using IAF model,the next is for the temporal external attention solely (TA-TR-I3D-IAF) , which consists of combining the temporal self attention by I3D and their fusion using IAF model. The last line is our STA-IAF, which dissociates the two attention modules in two separate streams and then combines them with I3D by the IAF fusion module.

Table 2. Comparison of Accuracy (%): STA-IAF vs. Various Model Configurations. without-TR designates without attention.

Configuration Method	NTU-Cross-Sub	NTU-Cross-View
I3D (without-TR)	85.5	87.3
ST-GCN-I3D-IAF	90.8	92.5
ST-GCN-TR-I3D-IAF	92.6	94.3
SA-TR-I3D-IAF	92.9	95.9
TA-TR-I3D-IAF	93.3	95.3
Dissociated STA-I3D-IAF	**94.8**	**97.1**

As displayed in Table 2, our STA-IAF model outperforms all other configurations by a significant margin. Unlike in STGCN-TR-I3D-IAF, the association

of the spatial and temporal attention modules in a single stream to guide RGB cues through 3D ConvNet decreases the classification performance (accuracy). This can be explained by the nature of the activities included in the NTU—RGB + D benchmark; In fact, it involves activities with significant motion variation (such as punching, slapping, kicking, etc.) that require the application of the temporal external attention module to help the model focus on specific temporal segments of the action sequence yielding to accurate action recognition. The dataset involves also human interactions with objects (such as playing with phone/tablet, put on a hat/cap, throw up cap/hat, etc.) that require the module of spatial self attention for the representation of the manipulated objects for the discrimination between actions. Therefore, the spatial self attention and the temporal external attention modules contribute in the improvement of the classification performance. As a consequence, we consider to dissociate the two attention modules instead of associating them in succession.

5 Conclusion

In this study, we introduced an innovative approach for human activity recognition, combining 3D skeletal data in spatial and temporal models based transformer network and RGB data as input to I3D network. Our approach surpasses the most advanced existing techniques when evaluated on the widely-used NTU—60 dataset. To establish an in-depth comprehension of semantics between multimodal characteristics, we incorporated the intermodal fusion module based attention (IAF). This module intelligently merges features using attention mechanisms, allowing us to capture the most discriminative features enabling to predict the action category. Our exhaustive experiments consistently demonstrate our superior performance compared to existing approaches. Our research addresses the challenges associated with disambiguating similar actions and recognizing actions presenting fine details (fine granularity actions) by introducing the attention mechanism in graphs networks. Additionally, we leverage the existing pretrained CNN backbone to improve the extraction of spatio-temporal characteristics in video sequences. However, it's worth noting that 3D CNN models, like I3D, are typically designed for short videos. Additional areas of focus for future work will involve expending our model's capabilities for handling long-term videos that involve complex actions, addressing the unique challenges posed by such scenarios.

Acknowledgments. This research work was supported by Abu Dhabi National Oil Company (ADNOC), Emirates NBD, Sharjah Electricity Water & Gas Authority (SEWA), Technology Innovation Institute (TII) and GSK as the sponsors of the 4^{th} Forum for Women in Research (QUWA): Sustaining Women's Empowerment in Research & Innovation at University of Sharjah.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
3. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
4. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Adv. Neural Inf. Process. Syst. **27** (2014)
5. Cheron, G., Laptev, I., Schmid, C.: P-CNN: pose-based CNN features for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 218–3226 (2017)
6. Ercolano, G., Riccio, D., Rossi, S.: Two deep approaches for ADL recognition: a multi-scale LSTM and a CNN-LSTM with a 3D matrix skeleton representation. In: 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 877–882 (2017)
7. Abu-El-Haija, S., et al.: YouTube-8M: a large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
8. Das, S., Koperski, M., Bremond, F., Francesca, G.: Action recognition based on a mixture of RGB and depth-based skeletons In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 IEEE (2017)
9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
10. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5323–5332 (2018)
11. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7912–7921 (2019)
12. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI Conference on Artificial Intelligence (2018)
13. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An End-to-End Spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
14. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. arXiv:1511.04119 (2015)
15. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. Adv. Neural Inf. Process. Syst. **30** (2017)
16. Long, X., Gan, C., De Melo, G., Wu, J., Liu, X., Wen, S.: Attention clusters: Purely attention based local feature integration for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7834–7843 (2018)
17. Baradel, F., Wolf, C., Mille, J., Taylor, G.W.: Glimpse clouds: human activity recognition from unstructured feature points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 469–478 (2018)

18. Zheng, Z., An, G., Wu, D., Ruan, Q.: Global and local knowledge-aware attention network for action recognition. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 334–347 (2020)
19. Guo, M.H., Liu, Z.N., Mu, T.J., Hu, S.M.: Beyond self-attention: External attention using two linear layers for visual tasks. arXiv preprint arXiv:2105.02358 (2021)
20. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12663, pp. 694–701. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68796-0_50
21. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 143–152 (2020)
22. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: multimodal transfer module for CNN fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13289–13299 (2020)
23. Baradel, F., Wolf, C., Mille, J.: Human action recognition: pose based attention draws focus to hands. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 604–613 (2017)
24. Baradel, F., Wolf, C., Mille, J.: Human activity recognition with pose-driven attention to RGB. In: BMVC 2018-29th British Machine Vision Conference, pp. 1–14 (2018)
25. Liu, G., Qian, J., Wen, F., Zhu, X., Ying, R., Liu, P.: Action recognition based on 3D skeleton and RGB frame fusion. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 258–264, IEEE (2019)
26. Baradel, C., Wolf, F., Mille, J., Taylor, G.W.: Glimpse clouds: human activity recognition from unstructured feature points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 469–478 (2018)
27. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035, IEEE (2019)
28. Shi, F., et al.: Star: sparse transformer-based action recognition. arXiv:2107.07089 (2021)
29. Li, C., Zhong, Q., Xie, D., Pu, S.: Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. arXiv:1804.06055 (2018)
30. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3595–3603, IEEE (2019)
31. Cho, S., Maqbool, M., Liu, F., Foroosh, H.: Self-attention network for skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 635–644, IEEE (2020)
32. Sun, Y., Shen, Y., Ma, L.: Msst-rt: multi-stream spatial-temporal relative transformer for skeleton-based action recognition. Sensors **21**(16), 5339 (2021)
33. Zhang, Z., Wang, Z., Zhuang, S., Huang, F.: Structure-feature fusion adaptive graph convolutional networks for skeleton-based action recognition. IEEE Access **8**, 228108–228117 (2020)

34. Liu, M., Yuan, J.: Recognizing human actions as the evolution of pose estimation maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1159–1168 (2018)
35. Das, S., et al.: Toyota smarthome: real-world activities of daily living. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 833–842, IEEE (2019)
36. Ahmad, T., Rizvi, S.T.H., Kanwal, N.: Transforming spatio-temporal self-attention using action embedding for skeleton-based action recognition. J. Vis. Commun. Image Represent. **95**, 103892 (2023)
37. Guddeti, R.M.R.: Human action recognition using multi-stream attention-based deep networks with heterogeneous data from overlapping sub-actions. Neural Comput. Appl. 1–17 (2024). https://doi.org/10.1007/s00521-024-09630-0