AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 56(1) (2024) 69-86
DOI: 10.22060/miscj.2024.23094.5357

Incorporating Transformer Networks and Joint Distance Images into Skeleton-driven Human Activity Recognition

Elham Shabaninia[1]*, Fatemeh Shafizadegan[2], Hossein Nezamabadi-pour[3], Ahmad R. Naghsh-Nilchi [2]

[1] Department of Applied Mathematics, Graduate University of Advanced Technology, Kerman, Iran
[2] Department of Artificial Intelligence, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran
[3] Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ABSTRACT: Skeleton-based action recognition has attracted significant attention in the field of computer vision. In recent years, Transformer networks have improved action recognition as a result of their ability to capture long-range dependencies and relationships in sequential data. In this context, a novel approach is proposed to enhance skeleton-based activity recognition by introducing Transformer self-attention alongside Convolutional Neural Network (CNN) architectures. The proposed method capitalizes on the 3D distances between pair-wise joints, utilizing this information to generate Joint Distance Images (JDIs) for each frame. These JDIs offer a relatively view-independent representation, allowing the model to discern intricate details of human actions. To further enhance the model's understanding of spatial features and relationships, the extracted JDIs from different frames are processed. They can be directly input into the Transformer network or first fed into a CNN, enabling the extraction of crucial spatial features. The obtained features, combined with positional embeddings, serve as input to a Transformer encoder, enabling the model to reconstruct the underlying structure of the action from the training data. Experimental results showcase the effectiveness of the proposed method, demonstrating performance comparable to other state-of-the-art transformer-based approaches on benchmark datasets such as NTU RGB+D and NTU RGB+D120. The incorporation of Transformer networks and Joint Distance Images presents a promising avenue for advancing the field of skeleton-based human action recognition, offering robust performance and improved generalization across diverse action datasets.

Review History:

Received: Apr. 08, 2024
Revised: Jul. 26, 2024
Accepted: Aug. 20, 2024
Available Online: Sep. 05, 2024

Keywords:

Human Activity Recognition
Joint Distance Image
Vision Transformer
Deep Learning

1- Introduction

Over the past few years, there has been notable enthusiasm in the domain of 3D skeleton-based human action recognition [1]. This heightened interest can be attributed primarily to the concurrent advancements in sensor technology and the accessibility of extensive datasets. Additionally, the noteworthy achievements of deep learning techniques, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success, particularly in tasks centered around image-based processing [2, 3].

In the realm of skeleton-based motion recognition using Convolutional Neural Networks (CNNs), skeletons are usually converted into images, where relevant information is reflected within image colour and texture. Undoubtedly, the characteristics of these transformed images play a crucial role in the effectiveness of convolutional networks. For this purpose, two approaches are common. The first approach aims to capture both spatial and temporal information from skeleton sequences and convert them into image properties [4, 5]. Typically, these methods map the 5D space of skeletons (comprising three spatial coordinates, time, and joint index) onto a collection of colour images. Unfortunately, this projection process is not lossless, resulting in the loss of certain information during the transformation. The second approach considers spatial and temporal information independently, subsequently merging the extracted features or decisions based on each data. While CNNs encode spatial information effectively, existing deep architectures often employ limited solutions to encode temporal information, such as 3D filters, precomputed motion features, and Recurrent Neural Networks (RNNs) [6].

The transformer represents a novel encoder-decoder architecture that leverages the attention mechanism to differentially assign importance to various components of the input data [7]. Unlike traditional recurrent neural networks (RNNs), transformers are adept at processing sequential input data without strict requirements for sequential order. By utilizing the attention mechanism, transformers offer contextual information for any position within the input sequence, enabling increased parallelization and reduced training time compared to RNNs. Transformers have demonstrated remarkable success in Natural Language Processing (NLP) [7] and have expanded their application domain to vision tasks [8, 9]. Video recognition presents a

*Corresponding author's email: e.shabaninia@kgut.ac.ir compelling case as the application of transformers. Videos can be considered as a sequence of images, akin to how language processing involves sequences of tokens [10].

This paper extends our previous work [11] by exploring the effects of using transformers in skeleton-based action recognition. In this approach, the 3D distances between pairs of joints are converted into Joint Distance Images (JDIs) for individual frames, and a vision transformer (ViT) is employed to encode temporal information within the sequence. Compared to [11], the primary novelty of this study lies in the comprehensive analysis and expansion of our findings, providing a more detailed understanding of the model's performance across diverse datasets and scenarios. In addition to employing various transformer-based networks, the proposed method demonstrates its capability in conjunction with different transformer architectures, such as the Vision Transformer. The key advancements in this version revolve around an in-depth exploration of the experimental outcomes, offering nuanced insights into the effectiveness of our proposed methodology.

The subsequent sections of this paper are organized as follows. Section 2 reviews related work. The proposed method is discussed in detail in section 3. Experimental results are provided in section 4. The paper concludes in section 5.

2- Related work

In vision-based human activity recognition, different modalities are used, such as RGB, depth, skeleton, and IR. While RGB data provide the shape, colour, and texture information, they are sensitive to viewpoint, illumination, and background variations. The depth modality is invariant to illumination and appearance changes, offering useful 3D structural information about the scene [2]. However, depth data are noisy and lack of colour and texture details. On the other hand, the positions of human joints in the skeletal data provide high-level information for motion recognition [2]. Skeletal data, on the other hand, provide information about the positions of human joints, which is crucial for motion recognition. Skeletal data require a low-dimensional space and are insensitive to motion speed, scale, and background variations. However, they do not provide information about objects in the scene for human-object interactions. The IR modality is suitable for dark environments. Due to the different characteristics of these modalities, different categories of methods are adopted in the literature. Among all these categories, skeletal joint-based techniques try to capture the spatio-temporal evolution of joints over time.

2- 1- Skeleton-based action recognition

The robustness of skeletal joint features to variations in camera location and subject appearance makes them an attractive choice for activity recognition algorithms. This invariance allows for the design of algorithms that can effectively handle different views and body sizes. As a result, there has been a significant research focus on utilizing skeletal joint information for activity recognition. While a comprehensive review of all existing methods is beyond the scope of this discussion, we will highlight some promising approaches. In a recent study [12], motion recognition techniques based on deep learning have been classified into two main categories: CNN-based methods and RNN/LSTM-based methods.

In CNN-based methods often a sequence of skeletons is first visualized as images. Li et al. [5] introduced Joint Distance Maps (JDM), a method that encodes the distances between skeleton joints for single or multiple subjects into images. These distance maps are relatively invariant to view variations. Liu et al. [4] proposed an improved skeleton visualization method for view-invariant human action recognition. This method utilizes a sequence-based view-invariant transform to normalize the spatial-temporal locations of skeleton joints, which are then visualized as a series of colour images. Subsequently, a multi-stream CNN fusion approach is employed for recognition. In [13], invariant features of translation, scale, and rotation are extracted for each body part of human skeleton sequences. Then these features are transformed into images to be fed to a CNN-based network. A skeleton image representation named SkeleMotion is introduced in [14], for input into CNNs. The suggested method captures the temporal dynamics by directly calculating the magnitude and orientation values of the skeletal joints. Skepxels (skeleton picture elements or skeleton pixels) are introduced in [15] as visual units to construct skeletal images directly processable by CNNs. This method organizes skeleton joints of a frame in a 2D grid to leverage 2D kernels in CNNs, while also capturing temporal evolution by combining Skepxels from multiple frames into a single image.

Graph-based approaches have gained considerable attention in recent years, complementing the conventional CNN methods [16, 17]. These approaches leverage the power of graph convolutional networks (GCN), which extend the concept of convolution from grid-based data to graph-based data. With GCNs, the iterative processing of graphs becomes possible, enabling the transformation of node features and their neighbouring nodes. Yan et al. [18] introduced spatial-temporal graph convolutional networks (ST-GCN) for the purpose of skeleton-based action recognition. The proposed method applies a series of spatial temporal graph convolutions to the skeleton sequences. By leveraging this technique, the authors effectively capture the spatiotemporal relationships present in the skeletal data. In [18], spatial temporal graph convolutional networks (ST-GCN) are proposed for skeleton-based action recognition, that perform a set of spatial temporal graph convolutions on the skeleton sequences.

In the second class of approaches, the sequential information of skeletal features across consecutive frames is utilized by employing RNNs or LSTMs as the time series to exploit the temporal evolution. Shahroudy et al. [19] introduced an architecture known as P-LSTM, which focuses on recognizing human actions while considering the distinct context of individual body parts. The P-LSTM model is designed to represent the output of its LSTM units as a combination of separate body parts, allowing for a more comprehensive understanding of the overall action being performed. This approach aims to leverage the temporal dynamics present in the data. Liu et al. [20] presented the Global Context-Aware Attention LSTM, which aims to enhance the attention mechanism within an action sequence by leveraging global context information. The proposed model effectively directs its focus towards the most informative joints during the sequence. Lee et al. [21] put forth a method for skeleton-based action recognition utilizing multiple Temporal Sliding LSTM (TS-LSTM) networks. The proposed model incorporates multiple TS-LSTM networks to capture temporal dependencies and enable accurate recognition of actions based on skeletal data. Zhang et al. [22] introduced two view adaptive neural networks, namely VA-RNN and VA-CNN, which offer an end-to-end solution for human action recognition. These networks are designed to dynamically adjust the observation viewpoints to the most suitable ones for action recognition. Zhang et al. [23] presented the Element-wise-Attention Gate (EleAttG) as a method to enhance the attentiveness capability of neurons within RNNs. This approach introduces an EleAttG module that can be seamlessly incorporated into an RNN block. Unlike traditional methods that modulate the input as a whole, EleAttG operates elementwise, allowing for content-adaptive modulation of the input.

2- 2- Transformers for action recognition

Vaswani et al. [7] introduced the transformer as a viable alternative to recurrent networks for sequence modelling. This model has emerged as the leading approach in natural language processing (NLP) due to its exceptional ability to capture long-range dependencies using the self-attention mechanism and its capacity to parallelize input processing. While initially developed for NLP, the transformer's self-attention mechanism has found widespread utility in various computer vision tasks, including image classification (in studies like ViT and DeiT [8, 24]), object detection [25], video instance segmentation [26], action recognition [27], thus showcasing its versatility beyond its original domain. However, there are still limited works for skeleton-based action recognition using transformers [28]. To address this gap, Plizzari et al. [29] introduced a Spatial-Temporal Transformer network (ST-TR) specifically designed to capture dependencies between joints by leveraging the transformer's self-attention operator. The ST-TR model incorporates a Spatial Self-Attention module (SSA) that focuses on understanding intra-frame interactions among various body parts. Additionally, a Temporal Self-Attention module (TSA) is employed to model inter-frame correlations, enabling the network to analyze the temporal dynamics of the action sequence. A transformer architecture called the relative transformer was introduced in [30]. This lightweight transformer is designed to establish connections between distant joints within a spatial framework, enabling efficient signal propagation. Additionally, it is utilized in the temporal dimension to effectively capture and model long-range interactions between distant frames. In [31], a novel unsupervised learning framework called the hierarchical transformer was presented which aimed to enhance skeleton-based human action recognition. This framework incorporates self-attention modules in a hierarchical manner to effectively capture the spatial and temporal structure within the skeleton sequences. Cheng et al. [32] introduced a transformer-based model known as the motion transformer, which aimed to effectively capture temporal dependencies through self-supervised pre-training on human action sequences. The proposed framework combines a transformer-based approach with a 2D convolutional network to learn spatial feature representations from joint distance images. The model then incorporates attention mechanisms to incorporate temporal information into the feature flow.

3- Material and method

In this section, a novel method is introduced for recognizing different human actions using skeletal data. This approach involves extracting spatial and temporal information from linear projections of flattened JDIs via a vision transformer. Additionally, similar to the approach described in [11], visual features can be extracted from JDIs for each frame using CNNs. The temporal information is then encoded using a transformer encoder (see Fig. 1).

3- 1- Joint distance image (JDI)

The input of the algorithm is a sequence of skeletons containing the 3D position of joints in successive frames. In each frame, the distances of skeleton joints related to single or multiple subjects are arranged into an image, called JDI. These distance images are relatively invariant to view variations. For a frame, containing one skeleton of joints, the JDI is an colour image where pixel (i, j) is the difference of the ith joint and the jth joint. i.e.

$$JDI_x(i,j) = x_i - x_j \tag{1}$$

$$JDI_y(i,j) = y_i - y_j \tag{2}$$

$$JDI_z(i,j) = z_i - z_j \tag{3}$$

where is the 3D position of the mth joint. The interesting property of these images is that the JDIs are skew-symmetric in each channel. i.e.

$$JDI_{x,y,z}{}^T = -JDI_{x,y,z} \tag{4}$$

When multiple skeletons are present in a scene, the joints are arranged for the first skeleton, then for the second one, etc. In this case, the JDI represents some clusters, where the number of clusters is equal to two times the number of

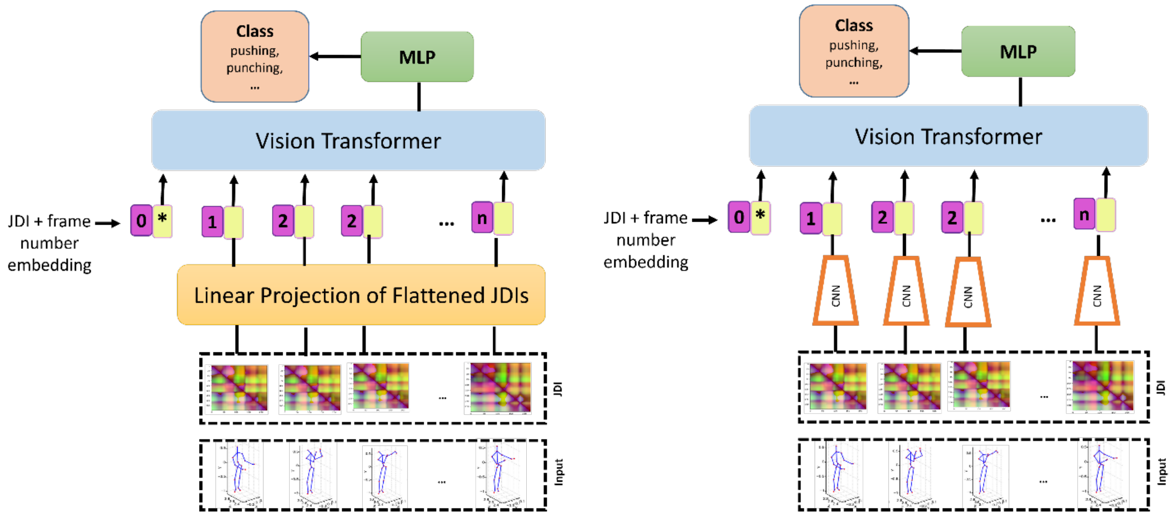


Fig. 1. The framework of the proposed method in this paper (left) and in [11] (right).

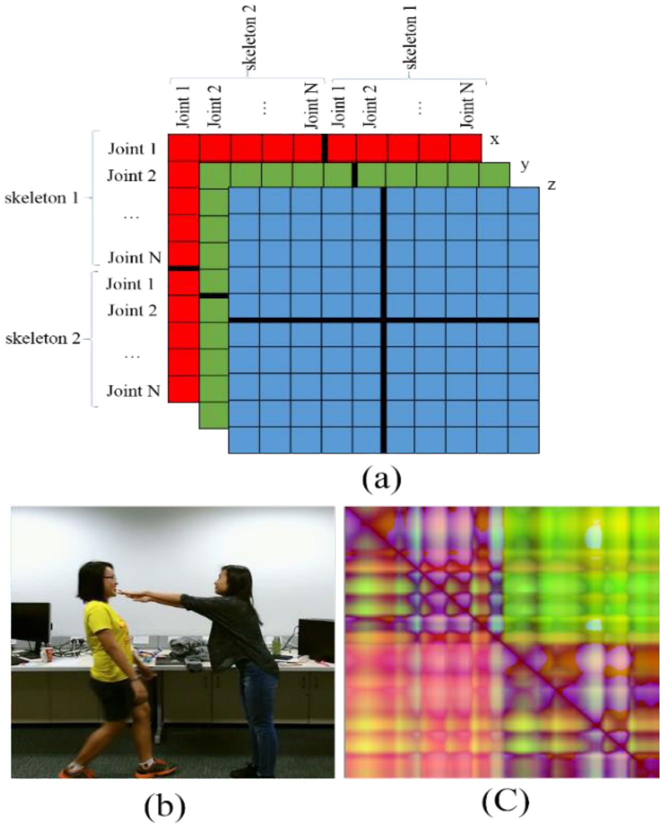


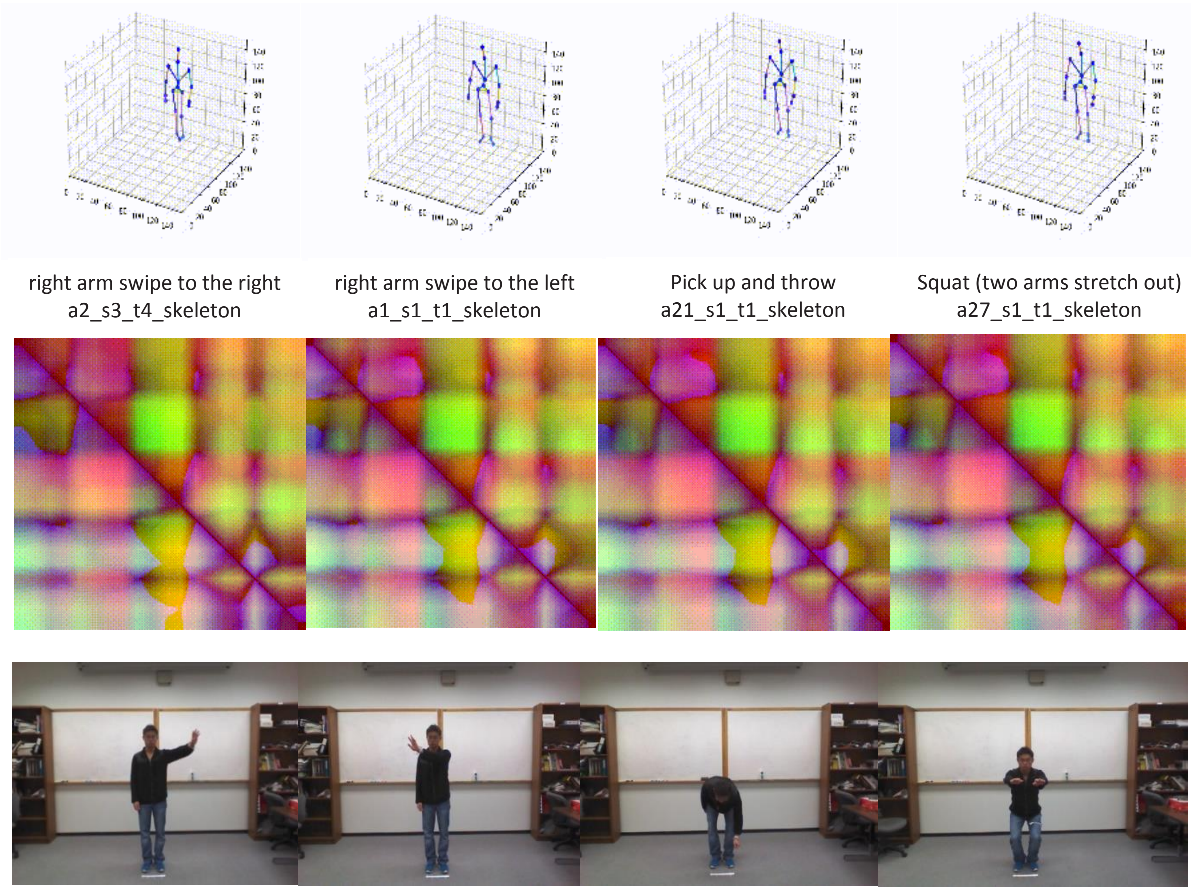
Fig. 2. (a) Constructing the JDI image when two skeletons are present in a frame. (b) A sample frame of action 'pushing' from NTU RGB+D dataset [19]. (c) The related JDI of the frame in (b). [11, 33]

skeletons (see Fig. 2).

JDMs are made for each frame of the sequence. Fig. 3 illustrates the average JDMs for two different actions. As this figure shows, these images are visually different. On the other hand, JDIs are relatively view-invariant. Fig. 4 shows an action ('walk around') from two different views. As this figure shows, although the 3D coordinates of joints are different, the average resulting JDIs are very similar.

3- 2- Visual features

The proposed method commences with spatial deep convolutional neural networks for extracting convolutional



right arm swipe to the right
a2_s3_t4_skeleton

right arm swipe to the left
a1_s1_t1_skeleton

Pick up and throw
a21_s1_t1_skeleton

Squat (two arms stretch out)
a27_s1_t1_skeleton

Fig. 3. The average JDIs for four different actions.

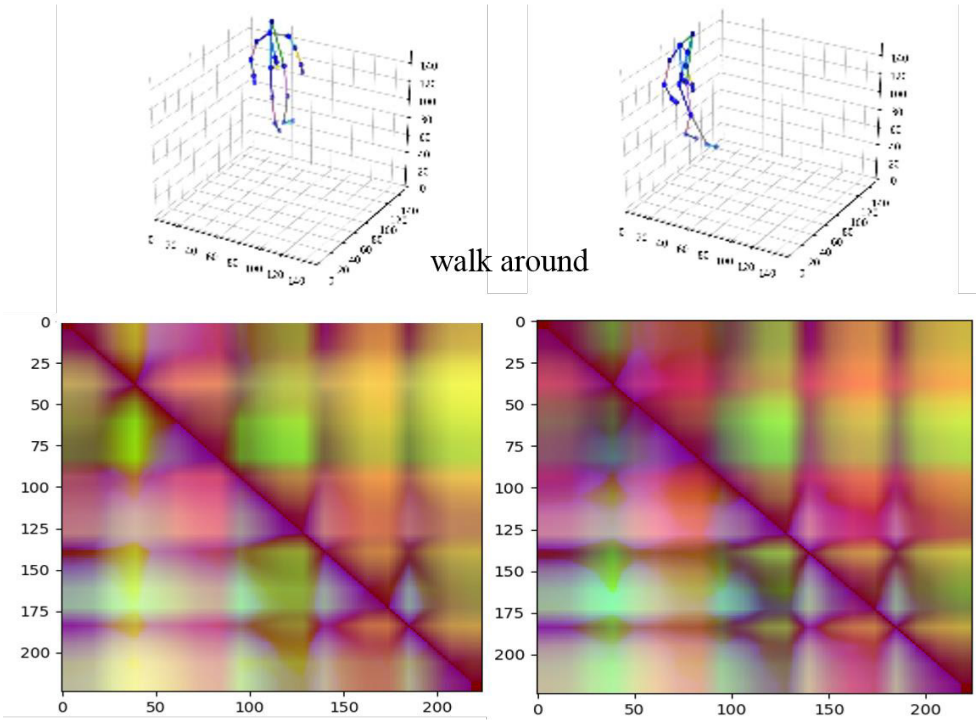


Fig. 4. The average JDI for action ‘walk around’ for two different views [11, 33].

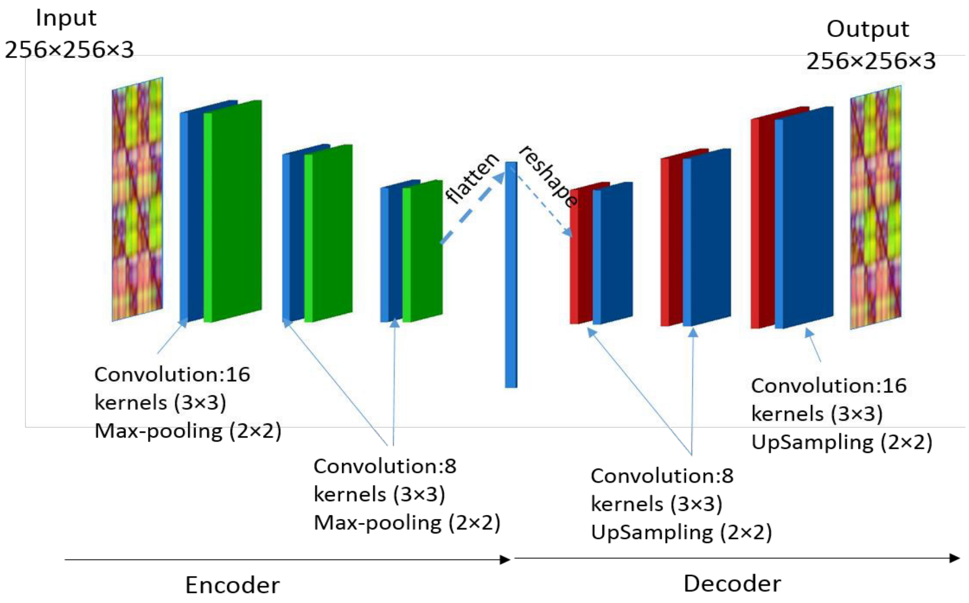


Fig. 5. The convolutional autoencoder trained on JDIs. The encoder part is used later for extracting spatial features [11].

feature maps. Here a pre-trained convolutional autoencoder is used for this purpose (see Fig. 5) instead of using existing backbones pre-trained on 2D natural images. That is because the JDIs are Essentially different from natural images. The network is trained on a set of collected JDIs and the encoder is frozen to extract an 8192-dimensional representation for each video frame. A spatial net (trained on 256×256×3 frame images) is designed for capturing appearance information. This single-frame architecture is based on 2D CNN model and extracts a feature vector for each frame.

3- 3- Sequence modelling

The transformer receives a 1D sequence of token embeddings as input. The spatial features x\in\mat extracted from CNN for all frames of a sequence are considered as JDI embeddings where N is the number of frames and D is the dimension of the flatted output latent vector of CNN. Frame number embeddings are added to the JDI embeddings to retain positional information. We use standard learnable 1D position embeddings for this purpose. A learnable embedding (similar to the [class] token of BERT) is added to the beginning of the sequence of embedded frames ($z_0^0 = x_{class}$).

$$z_0 = \left[x_{class};\ x_p^1;\ x_p^2;\ \dots;\ x_p^N\ \right] + E_{pos}$$
$$E_{pos} \in \mathbb{R}^{(N+1)\times D} \quad (5)$$

After propagating the sequence through the transformer layers consisting of alternating layers of multiheaded self-attention (MSA),

$$z_l' = MSA\big(LN(z_{l-1})\big) + z_{l-1}, \quad l = 1 \dots L \quad (6)$$

$$z_l = MLP\big(LN(z_l')\big) + z_l', \quad l = 1 \dots L \quad (7)$$

the final state of the features related to this classification token () is used as the final representation of the video and is applied to the given classification task head. The classification token is processed with an MLP (Multi-Layer Perceptron) head to provide a final predicted category. The MLP head contains two linear layers with a GELU non-linearity and Dropout between them. The input token representation is first processed with a Layer normalization (LN).

4- Experimental results

In this section, the experiments are performed to validate the efficiency of the proposed method on NTU RGB+D [19] (NTU60) and NTU RGB+D120 [34] (NTU120) datasets. The details are provided in the following.

NTU60 and NTU120 are large-scale datasets for RGB+D human action recognition. An example of NTU60 is shown in Fig. 6 with RGB and corresponding skeleton modality

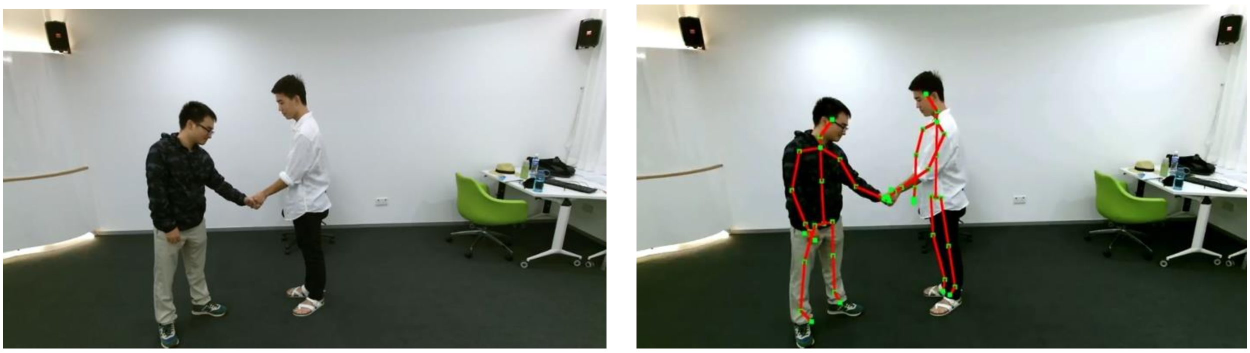


Fig. 6. Hands shaking from NTU60. (left) RGB modality and (right) skeleton mapped on RGB.

mapped on the RGB data. These datasets contain respectively 60 and 120 different action classes including daily, mutual, and health-related actions. The 3D data is captured by Kinect v2 cameras. They consist of different views (front view, two side views, and left/right 45-degree views). Large intra-class variations and different views make these datasets very challenging.

NTU60 consists of more than 56 thousand video samples and 4 million frames collected from 40 distinct subjects (aged between 10 and 35). Identical to [19], evaluations on NTU60 are performed by both cross-subject (CS) and cross-view (CV) evaluation protocols. In the cross-subject evaluation, samples of subjects 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38 are used as training and samples of the residual subjects are kept for testing. In the cross-view evaluation, samples taken by cameras 2 and 3 are used as training, while samples from camera 1 are preserved for testing. Through the experiments, for each action, 9 frames are chosen randomly from the entire video.

Cross-subject (CS) and cross-setup (CSet) evaluation protocols are performed on the NTU120 dataset, which contains almost 115000 video samples recorded from 106 subjects performing actions. For the cross-subject evaluation, 63360 samples of 53 subjects are used in training and 51120 samples of 53 subjects are used for testing. There are 54720 samples with 16 different setups used in training and 59760 samples with 16 other setups in the test [33].

4- 1- Cross-subject results

Cross-subject in the context of action recognition experiments typically refers to the performance of a model when it is trained on data from one set of subjects (participants) and tested on a different set of subjects. This approach is essential to assess the generalizability of a model across different individuals, ensuring that the model can recognize and classify actions accurately for people who were not part of the training dataset.

For example, in a scenario where you collect data from multiple individuals performing various actions, the cross-subject evaluation involves training the action recognition model on a subset of these individuals and evaluating its performance on the actions performed by individuals who were not part of the training set. This simulates how well the model can generalize to new subjects or users.

The cross-subject experimental setup helps researchers understand whether the learned patterns and features are specific to the individuals in the training set or if the model can effectively generalize its knowledge to new subjects. It provides valuable insights into the model's robustness and applicability in real-world scenarios where the users or actors might vary. Table 1 shows the cross-subject results of the proposed method on NTU60 and NTU120 datasets. As this table shows compared with some existing approaches, the proposed method demonstrates better performance in cross-subject evaluation on both datasets. These results highlight the robustness and applicability of the proposed method in scenarios where users or actors may vary. However, the results from two recent approaches [35, 36] show superior results for human action recognition.

The confusion matrix of JDI+ViT method on the NTU60 and NTU120 using CS evaluation protocol is presented in Fig. 7 and Fig. 8, respectively. Table 2 demonstrates examples of qualitative results on “drop” action in NTU60 using CS evaluation protocol.

4- 2- Cross-view results

Cross-view results in the context of action recognition experiments pertain to evaluating the performance of a model when it is trained on data from one viewpoint or camera angle and tested on a different viewpoint. This is particularly relevant in scenarios where actions are captured from multiple perspectives, such as different camera placements or angles.

For example, in a surveillance system or a setting with multiple cameras, cross-view evaluation involves training the

Table 1. Cross-subject accuracy of methods on NTU60 and NTU120 datasets while top-5 accuracies are shown in parentheses.

Method	modality	NTU60	NTU120
Lie Group [37]	Skeleton	50.08	-
HBRNN [38]	Skeleton	59.07	-
1 Layer P-LSTM [19]	Skeleton	62.05	-
2 Layer P-LSTM [19]	Skeleton	62.93	-
SkateFormer [35]	Skeleton	93.5	89.8
STEP-CATFormer [36]	skeleton	93.2	90.0
JDI + VIVIT [39]	Skeleton	45.94 (79.35)	34.54 (65.54)
JDI+ CNN+ViT [11]	Skeleton	61.80 (89.03)	51.88 (82.01)
JDI+ViT (ours)	Skeleton	67.36 (91.39)	58.07 (85.25)

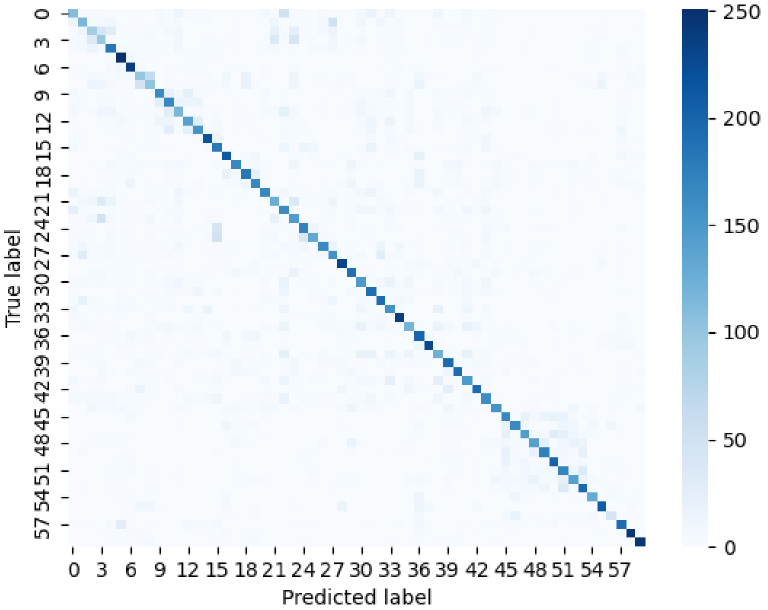


Fig. 7. Confusion matrix on NTU60 with CS evaluation.

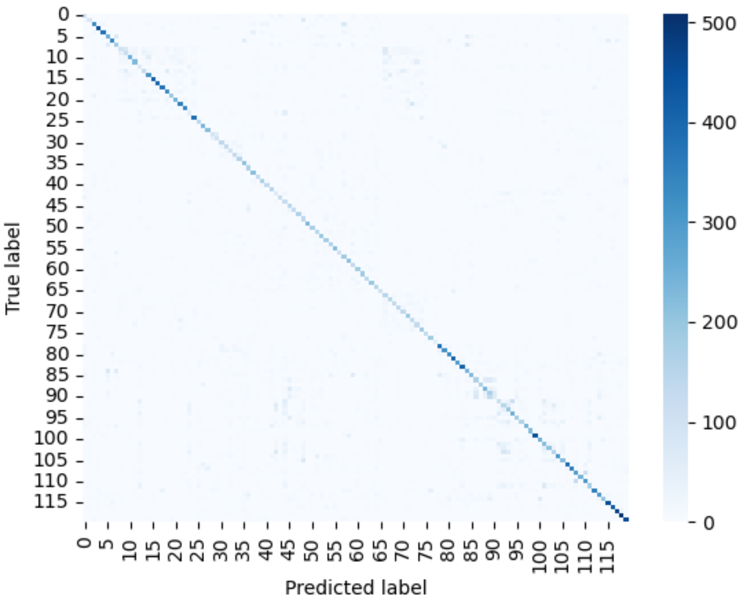


Fig. 8. Confusion matrix on NTU120 on CS evaluation.

Table 2. Examples of qualitative results on “drop” action in NTU60 with CS evaluation protocol.

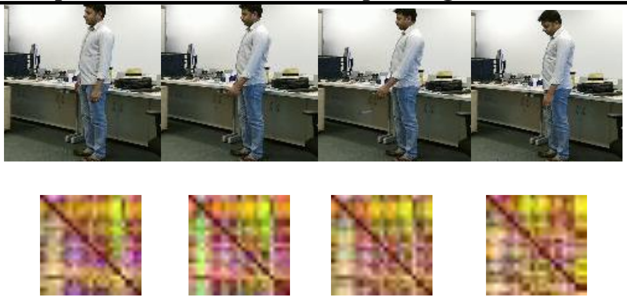
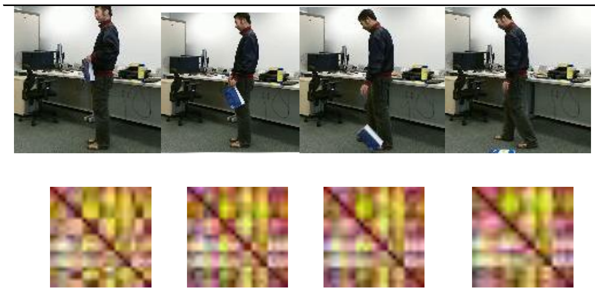
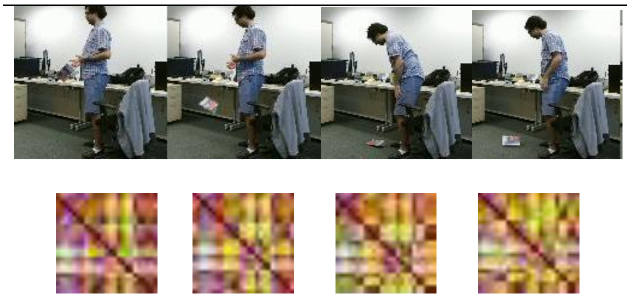
Sample RGB frames and corresponding JDIs	Predicted label	File ID
	Drop	S004C002P003R001A005
	Drop	S004C002P020R001A005
	Chest pain	S004C002P007R001A005

Table 3. Cross-view accuracy of methods on NTU60 dataset while top-5 accuracies are shown in parentheses.

Method	Modality	Accuracy
Lie Group [37]	Skeleton	52.76
HBRNN [38]	Skeleton	63.97
1 Layer P-LSTM [19]	Skeleton	69.40
2 Layer P-LSTM [19]	Skeleton	70.27
SkateFormer [35]	Skeleton	97.8
STEP-CATFormer [36]	Skeleton	97.3
JDI + VIVIT [39]	Skeleton	47.71 (81.47)
JDI+ CNN+ViT [11]	Skeleton	62.96 (90.23)
JDI+ViT (ours)	Skeleton	70.81 (93.70)

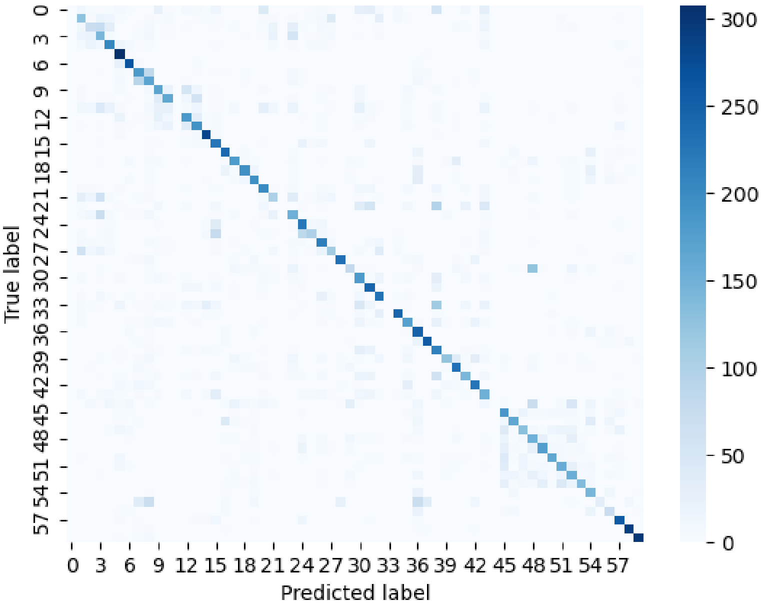


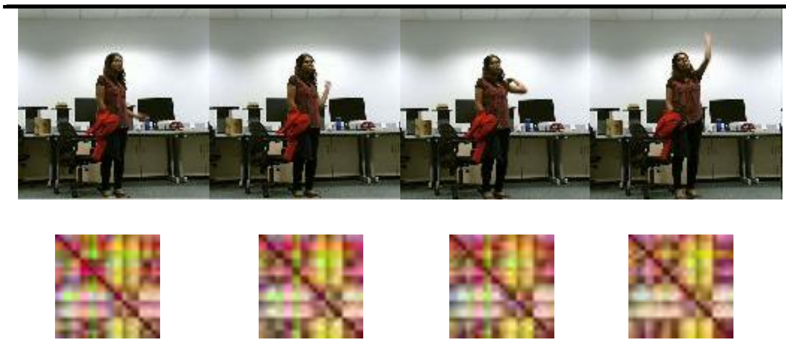
Fig. 9. Confusion matrix on NTU60 with CV evaluation.

action recognition model on data recorded from one viewpoint and then assessing its ability to accurately recognize actions when presented with data captured from a different viewpoint. This helps researchers and practitioners understand how well the model generalizes its learned features and patterns across different viewing angles.

Reporting cross-view results is crucial because it reflects the model's capability to recognize actions regardless of the specific camera perspective. In real-world applications, surveillance cameras or sensors may be distributed across various locations, each with its own viewpoint. Therefore, a robust action recognition model should be able to handle diverse viewing conditions and still accurately identify actions. Table 3 shows the cross-view results of the proposed method on NTU60 dataset. As this table shows compared with some other existing approaches, the proposed method achieves suitable results in recognizing actions across different camera perspectives. This highlights the model's capability to handle diverse viewing conditions and accurately identify actions. Although again, the results from [35, 36] show superior results for human action recognition.

The confusion matrix of JDI+ViT method on the NTU60 using CV evaluation protocol, is presented in Fig. 9. Table 4 demonstrates examples of qualitative results on "throw" action in NTU60 using CV evaluation protocol.

Table 4. Examples of qualitative results on "throw" action in NTU60 with CV evaluation.

Sample RGB frames and corresponding JDIs	Predicted label	File ID
	Throw	S007C001P016R002A007
	Throw	S010C001P016R002A007
	Throw	S015C001P016R001A007

4- 3- Cross setup

Combining these two evaluations, the "cross setup" in the NTU120 dataset involves training the model on a specific subset of subjects and views and testing it on data recorded from subjects and views not encountered during the training phase. This comprehensive evaluation helps researchers understand how well a model can generalize to new individuals and different camera angles, which is crucial for assessing its real-world applicability. Table 5 shows the cross-setup results of the proposed method on NTU120 dataset. These results validate the relative model's effectiveness in handling unseen subjects and camera setups, highlighting its potential for practical applications in real-world scenarios.

The confusion matrix of JDI+ViT method on the NTU120 using CSet evaluation protocol, is presented in Fig. 10. Table 6 shows examples of qualitative results on "run on the spot" action in NTU120 using CSet evaluation protocol.

In addition, the proposed method could also be employed with other transformer networks. To demonstrate its capability across different networks, various architectures are utilized instead of the Vision Transformer, as shown in Table 7. The results indicate that the ViT network yields the best performance.

As these tables show, the competitive results are obtained by the proposed method compared with other methods using transformers. These results show the ability of JDIs to capture view-invariant spatial features and the capabilities of vision transformers in capturing long-range temporal features of skeletal data.

In Table 8, examples of qualitative results of the proposed method are presented for different types of daily, medical conditions, and mutual actions. These results showcase how

Table 5. Cross-setup accuracy of methods on NTU120 dataset.

Method	modality	Accuracy
Lie Group [37]	Skeleton	-
HBRNN [38]	Skeleton	-
1 Layer P-LSTM [19]	Skeleton	-
2 Layer P-LSTM [19]	Skeleton	-
SkateFormer [35]	Skeleton	91.4
STEP-CATFormer [36]	Skeleton	91.2
JDI + VIVIT [39]	Skeleton	35.58 (66.81)
JDI+ CNN+ViT (ours)	Skeleton	50.12 (80.16)
JDI+ViT (ours)	Skeleton	55.68 (83.06)

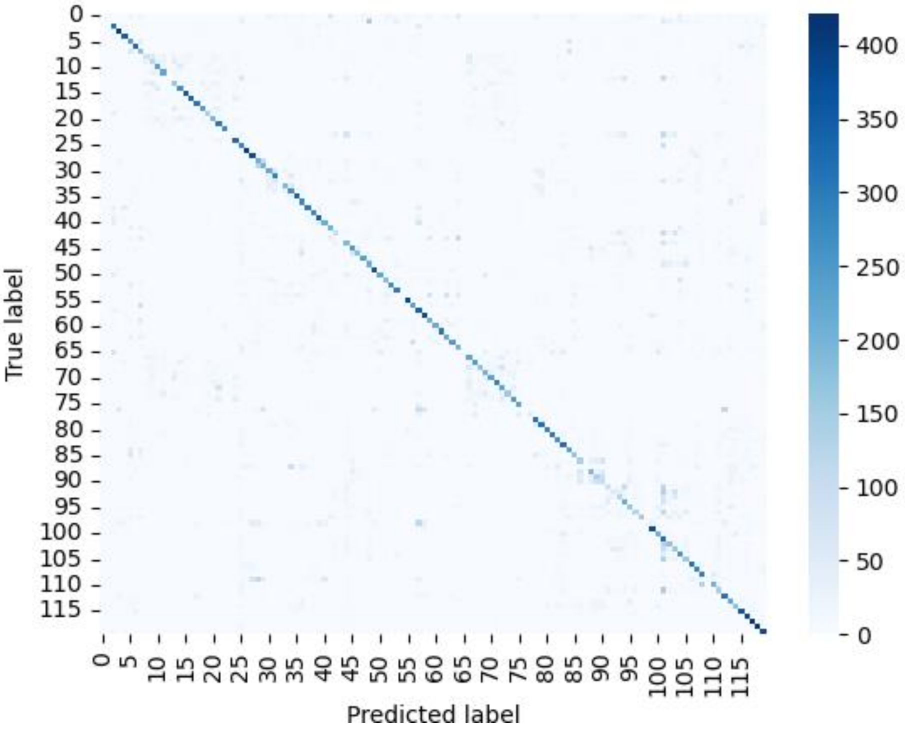


Fig. 10. Confusion matrix on NTU120 with CSet evaluation protocol.

Table 6. Examples of qualitative results on ‘run on the spot’ action in NTU120 with CSet evaluation protocol.

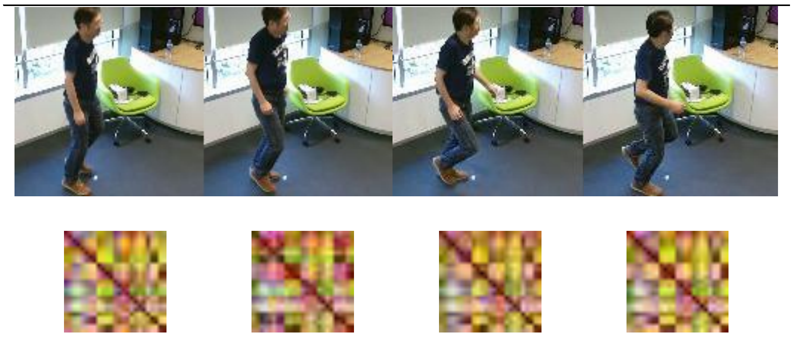
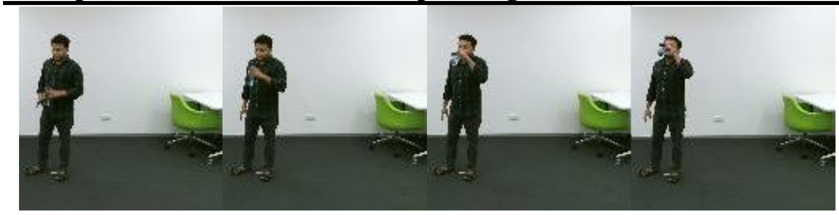
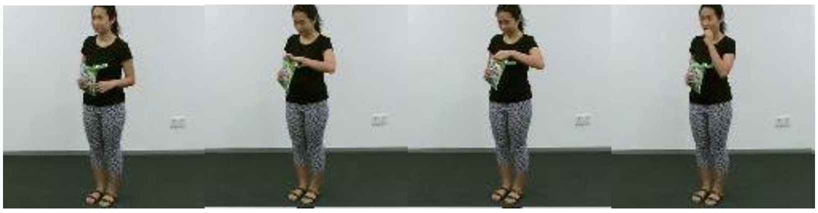
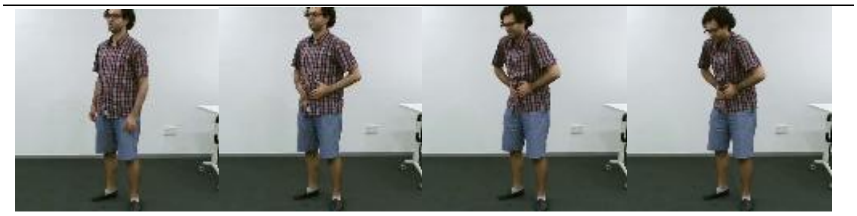
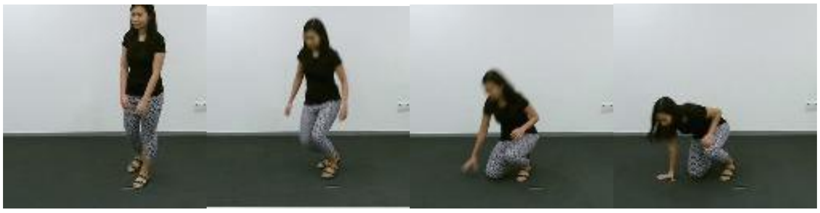
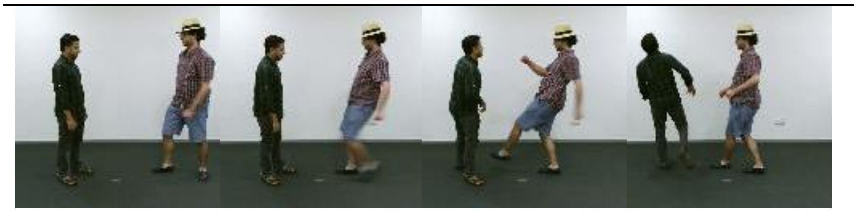
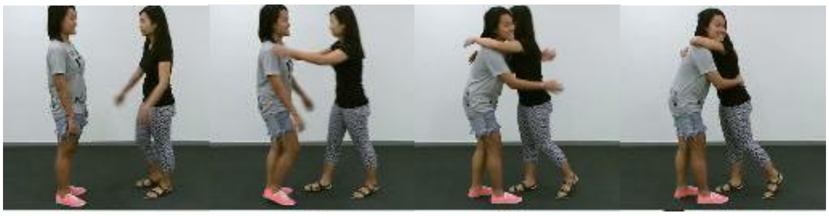
Sample RGB frames and corresponding JDIs	Predicted label	File ID
	Run on the spot	S032C002P104R002A099
	Run on the spot	S032C003P104R001A099
	Run on the spot	S032C003P104R002A099

Table 7. Results of using different transformer-based networks on NTU60 with cross-subject protocol (top-5 accuracies).

Method	modality	NTU60-cs
JDI + VIVIT [39]	Skeleton	79.35
JDI + VideoMAE [40]	Skeleton	80.11
JDI + Cait [41]	Skeleton	57.64
JDI + Deit [24]	Skeleton	21.75
JDI+ CNN +ViT [11]	Skeleton	89.03
JDI+ViT (ours)	Skeleton	**91.39**

Table 8. Examples of qualitative results on daily, medical condition, and mutual actions in NTU120.

Sample RGB frames and corresponding JDIs	Predicted label	Action type
	Drink water	**Daily**
	Eat meal	
	Chest pain	**Medical condition**
	Falling down	
	Kicking	**Mutual**
	Hugging	

the method performs in recognizing different actions based on skeletal data. The qualitative analysis provides insights into the robustness and generalizability of the proposed approach across a diverse range of scenarios and contexts.

5- Conclusion

This paper proposes a novel approach for skeletal action recognition, leveraging a transformer-based model. The method extracts features from JDIs at individual frames and incorporates temporal information by applying attention mechanisms to the resulting features. Experimental results demonstrate the effectiveness of the proposed method in capturing spatio-temporal features. While transformers have become the preferred model for NLP tasks, replacing RNN models like LSTM, their application in skeleton-based human action recognition is still in its early stages. Transformers are typically large and computationally intensive compared to CNN models. Therefore, developing high-performance transformer models with reduced resource requirements remains an ongoing challenge.

Acknowledgements

This work was partially supported by the Iran National Science Foundation (ISNF) and Shahid Bahonar University of Kerman, Grant No. 98006291.

Declarations

Ethical Approval
this declaration is “not applicable”.

Funding

This work was partially supported by the Iran National Science Foundation (ISNF) and Shahid Bahonar University of Kerman, GrantNo. 98006291.

Availability of data and materials

The datasets NTU RGB+D and NTU RGB+D 120 analyzed during the current study are available in https://rose1.ntu.edu.sg/dataset/actionRecognition/

References

[1] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, “Sensor-based and vision-based human activity recognition: A comprehensive survey,” Pattern Recognition, vol. 108, p. 107561, 2020.

[2] F. Shafizadegan, A. R. Naghsh-Nilchi, and E. Shabaninia, “Multimodal vision-based human action recognition using deep learning: A review,” Accepted in Artificial Intelligence Review, 2024.

[3] N. Imanpour, A. R. Naghsh-Nilchi, A. Monadjemi, H. Karshenas, K. Nasrollahi, and T. B. Moeslund, “Memory and time-efficient dense network for single-image super-resolution,” IET Signal Processing, vol. 15, no. 2, pp. 141-152, 2021.

[4] M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” Pattern Recognition, vol. 68, pp. 346-362, 2017.

[5] C. Li, Y. Hou, P. Wang, and W. Li, “Joint distance maps based action recognition with convolutional neural networks,” IEEE Signal Processing Letters, vol. 24, no. 5, pp. 624-628, 2017.

[6] Y. Kong and Y. Fu, “Human action recognition and prediction: A survey,” International Journal of Computer Vision, vol. 130, no. 5, pp. 1366-1401, 2022.

[7] A. Vaswani et al., “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998-6008.

[8] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.

[9] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” ACM computing surveys (CSUR), vol. 54, no. 10s, pp. 1-41, 2022.

[10] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, “Video transformer network,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3163-3172.

[11] E. Shabaninia and H. Nezamabadi-pour, “Skeleton-based human action recognition using joint distance images and vision transformers,” presented at the ICCE, 2023.

[12] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, “RGB-D-based human motion recognition with deep learning: A survey,” Computer Vision and Image Understanding, vol. 171, pp. 118-139, 2018.

[13] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, “Skeletonnet: Mining deep part features for 3-d action recognition,” IEEE signal processing letters, vol. 24, no. 6, pp. 731-735, 2017.

[14] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz, “SkeleMotion: A New Representation of Skeleton Joint Sequences Based on Motion Information for 3D Action Recognition,” in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, pp. 1-8: IEEE.

[15] J. Liu, N. Akhtar, and A. Mian, “Skepxels: Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition,” in CVPR Workshops, 2019.

[16] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, “Skeleton-based action recognition with shift graph convolutional network,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183-192.

[17] Y. Li, Z. He, X. Ye, Z. He, and K. Han, “Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition,” EURASIP Journal on Image and Video Processing, vol. 2019, no. 1, p. 78, 2019.

[18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Thirty-second AAAI conference on artificial intelligence, 2018.

[19] A. Shahroudy, J. Liu, T.-T. Ng, and G. 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, 2016, pp. 1010-1019.

[20] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 1586-1599, 2018.

[21] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 1012-1020.

[22] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 8, pp. 1963-1978, 2019.

[23] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks," IEEE Transactions on Image Processing, vol. 29, pp. 1061-1073, 2019.

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International Conference on Machine Learning, 2021, pp. 10347-10357: PMLR.

[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European Conference on Computer Vision, 2020, pp. 213-229: Springer.

[26] Y. Wang et al., "End-to-end video instance segmentation with transformers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8741-8750.

[27] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Trear: Transformer-based rgb-d egocentric action recognition," IEEE Transactions on Cognitive and Developmental Systems, 2021.

[28] E. Shabaninia, H. Nezamabadi-pour, and F. Shafizadegan, "Multimodal action recognition: a comprehensive survey on temporal modeling," Multimedia Tools and Applications, pp. 1-51, 2023.

[29] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," in International Conference on Pattern Recognition, 2021, pp. 694-701: Springer.

[30] Y. Sun, Y. Shen, and L. Ma, "MSST-RT: Multi-Stream Spatial-Temporal Relative Transformer for Skeleton-Based Action Recognition," Sensors, vol. 21, no. 16, p. 5339, 2021.

[31] Y.-B. Cheng, X. Chen, J. Chen, P. Wei, D. Zhang, and L. Lin, "Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition," in 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6: IEEE.

[32] Y.-B. Cheng, X. Chen, D. Zhang, and L. Lin, "Motion-transformer: self-supervised pre-training for skeleton-based action recognition," in Proceedings of the 2nd ACM International Conference on Multimedia in Asia, 2021, pp. 1-6.

[33] F. Shafizadegan, A. R. Naghsh-Nilchi, and E. Shabaninia, "Hybrid Embedding for Few-Frames Action Recognition Using Vision Transformers," Under review in internationl journal of multimedia information retrieval.

[34] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 10, pp. 2684-2701, 2019.

[35] J. Do and M. Kim, "SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition," arXiv preprint arXiv:2403.09508, 2024.

[36] B. L. N. Huu and T. Matsui, "Step catformer: Spatial-temporal effective body-part cross attention transformer for skeleton-based action recognition," 2022.

[37] R. Vemulapalli, F. Arrate, and R. 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, 2014, pp. 588-595.

[38] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1110-1118.

[39] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6836-6846.

[40] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," Advances in neural information processing systems, vol. 35, pp. 10078-10093, 2022.

[41] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 32-42.

HOW TO CITE THIS ARTICLE

E. Shabaninia, F. Shafizadegan, H Nezamabadi-pour, A. R. Naghsh-Nilchi, Incorporating Transformer Networks and Joint Distance Images into Skeleton-driven Human Activity Recognition, AUT J. Model. Simul., 56(1) (2024) 69-86.

DOI: 10.22060/miscj.2024.23094.5357