International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

An Improvements of Deep Learner Based Human Activity Recognition with the Aid of Graph Convolution Features

N. Srilakshmi[1], N.Radha[2]
[1]Ph.D., Scholar, Department of Computer Science,
PSGR Krishnammal College for Women,
Coimbatore- 641004 , Tamilnadu, India
srilakshmiphd123@gmail.com
[2]Associate Professor, Department of Computer Science,
PSGR Krishnammal College for Women,
Coimbatore - 641004 , Tamilnadu, India
radha@psgrkcw.ac.in

Abstract— Many researchers are now focusing on Human Action Recognition (HAR), which is based on various deep-learning features related to body joints and their trajectories from videos. Among many schemes, Joints and Trajectory-pooled 3D-Deep Geometric Positional Attention-based Hierarchical Bidirectional Recurrent convolutional Descriptors (JTDGPAHBRD) can provide a video descriptor by learning geometric features and trajectories of the body joints. But the spatial-temporal dynamics of the different geometric features of the skeleton structure were not explored deeper. To solve this problem, this article develops the Graph Convolutional Network (GCN) in addition to the JTDGPAHBRD to create a video descriptor for HAR. The GCN can obtain complementary information, such as higher-level spatial-temporal features, between consecutive frames for enhancing end-to-end learning. In addition, to improve feature representation ability, a search space with several adaptive graph components is created. Then, a sampling and computation-effective evolution scheme are applied to explore this space. Moreover, the resultant GCN provides the temporal dynamics of the skeleton pattern, which are fused with the geometric features of the skeleton body joints and trajectory coordinates from the JTDGPAHBRD to create a more effective video descriptor for HAR. Finally, extensive experiments show that the JTDGPAHBRD-GCN model outperforms the existing HAR models on the Penn Action Dataset (PAD).

Keywords- Human action recognition, JTDGPAHBRD, Skeleton structure, Geometric relations, Graph convolutional network, Spatial-temporal features, Temporal dynamics.

I. INTRODUCTION

Human Action Recognition (HAR) which automatically recognize and categorizes individual activities in videos, is one of the foremost essential active fields in artificial intelligence [1-3]. Typically, it holds great significance in the areas of audio-visual analysis [4-5], virtual reality [6-8], smart person-machine interactions [9-10], and so on [11-12]. Human actions can be recognized in several modalities, like RGB and skeletons. The skeleton structures express compressed details regarding an individual motion that could offer a powerful and vigorous exemplification to describe individual activities.

The skeleton information containing 3D coordinates of major joints in the individual body will be effortlessly captured thanks to advancements in depth sensors. So, skeleton-based HAR schemes are widely used nowadays [13]. Many conventional skeleton-based HAR schemes formulate the presence and the temporal dynamics of joints with handcrafted features like the relative positions between joints, the angles between limbs, and the surfaces covered by the human body. Skeleton-based features, on the other hand, are local characteristics containing coordinates of joints and their high-level correlations. So, those schemes are not appropriate for model and differentiating activities with related poses, movements, and person-machine interfaces [14]. They also rely heavily on skeleton prediction, and improper body joint discovery was not avoided. To tackle these issues, Convolutional Neural Networks (CNNs) are deployed in video-based HAR, which captures local-to-global features from both RGB images and depth.

From this perspective, a Joint and Trajectory-pooled 3D convolutional Descriptor (JTDD) scheme [15] was developed to extract and merge the body joint coordinates and their trajectories in the 2-stream Convolutional 3D (C3D) network for HAR. On the other hand, this scheme neglects the relevant spatial variations among various actions due to the use of max-min pooling in the C3D network, which was very flexible to spatially smooth over the nearby kernels. Therefore, the JTDPABRD scheme [16] was designed to substitute the max-min pooling by the Positional Attention-based Bidirectional

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

Recurrent Neural Network (PABRNN) for feature representation. But the BRNN has more parameters, which results in a vanishing gradient problem and a lack of learning long-range joint relations between actions.

So, the JTDPAHBRD scheme was developed by applying the PAHBRNN, which divides the feature maps related to the human skeleton in every clip into different parts depending on the body structure [17]. Those part-based features were hierarchically learned by the separate PABRNNs to obtain and fuse the long-range spatiotemporal information related to the different body parts. In contrast, these variants of JTDD for HAR were based on the video descriptor, which was created by merging only the joint and trajectory coordinates of different body parts at each time step. It was essential to obtain the geometric correlation among body joints for generating more meaningful descriptors. Since the trajectories of the body's joints only express gesture information and do not define contour or geometrical relations.

Accordingly, the skeleton graph was considered to determine the different types of geometries, like joints, edges, and surfaces, along with the trajectories of body joints [18]. This data was sent to the C3D network, which includes the novel View Conversion (VC) layer and the Temporal Dropout (TD) layer in the attention and feature streams, respectively, to learn the temporal dynamics of various geometries. Also, the PAHBRNN was used to obtain the final feature representation. After that, the outcomes of the two streams were multiplied by the bilinear pooling followed by the fully connected layer. The entire net was trained by using the softmax loss function to construct the video descriptor of a given frame, which was classified by the SVM classifier to classify human activities. In contrast, the spatial-temporal dynamics of the different geometric features of the skeleton structure were not explored deeper.

Hence in this paper, the GCN is incorporated with the JTDGPAHBRD scheme for HAR. The GCN learns complementary information between consecutive frames, such as higher-level spatial-temporal features, to improve end-to-end learning and generate video descriptors for a specific video sequence. A search space with several adaptive graph components is created to improve feature representation ability. After that, a sampling and computation-effective evolution scheme are applied to explore this space. So, the resultant GCN provides the temporal dynamics of the skeleton pattern, which are fused with the geometric features of the skeleton body joints and trajectory coordinates from the JTDGPAHBRD to create a more effective video descriptor for HAR. Such obtained descriptors are later classified by the SVM algorithm to recognize a variety of human actions. Thus, this model increases the accuracy of recognizing different kinds of human actions effectively.

The remaining sections are prepared as the following: Section II reviews the works associated with this study. Section III describes the JTDGPAHBRD-GCN and Section IV validates its recognition rate. Section V outlines the findings of this study and suggests future enhancement.

II. LITEARTURE SURVEY

Wan et al. [19] developed a 2-stream CNN for extracting long-short-term spatiotemporal characteristics, which were merged and classified by the linear SVM for HAR. However, accuracy was low on a few more similar action classes, and feature extraction was difficult for videos with complex backgrounds. Li et al. [20] developed a new Temporal Segment Connection Network (TSCN) for recognizing human actions. The forget-gate connection unit was used to extract and fuse deep features from multiple sampling groups, which provides a more global feature representation for actions. An adaptive weighting unit was used to learn multiple weights for various sampling groups. However, this model was inefficient because it required more memory and only benefited more heterogeneous databases.

Ren et al. [21] developed a new Segment of Cooperative Convolutional Networks (SC-ConvNets). Initially, segmented rank pooling was used to map the whole RGB-D frames into photos, which were fed to the ConvNets to define the spatiotemporal data. Then, a mutual optimization error value was applied to train complementary characteristics for multimodal HAR. But it did not simultaneously train the discriminatory characteristics of 3D multimodal data.

Hao et al. [22] presented a Hyper-Graph Neural network (Hyper-GNN) to extract spatiotemporal data and high-order correlations for HAR. First, the underlying skeleton graph was extended to define the high-level correlations by the hyperedge structure, and the convolution process was applied to the hypergraph. Then, the spatial co-occurrence trait was induced and the time-based correlation was added to the upgraded residual unit to capture wealthier characteristics. Moreover, a dynamic fusion of the 3-stream model was used to merge different features and recognize actions. But the accuracy was degraded while increasing the number of hyperedges, which may produce noise.

Li et al. [23] designed a new Symbiotic GNN (Sybio-GNN) to utilize graph-based operations for learning action patterns that simultaneously handle HAR and motion prediction. It comprises a support, an activity detection head, and a movement estimation head. For the support, multi-branch, multi-scale GCN was applied to capture spatiotemporal characteristics based on joint-scale and part-scale graphs. Also, twin bone-based graphs and nets were used to train complementary characteristics for HAR. But it used only the long-range joint relations, whereas the short-range joint relations and temporal features were not learned.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

Cha et al. [24] designed 3D interconnects of the individual bodies from RGB videos for HAR. The transformer structure was used to obtain a useful skeletal interpretation from the rebuilt 3D interconnects. However, due to inefficient memory, cooperatively training the bone and skeleton joint positions proved difficult. Yadav et al. [25] developed the Convolutional Long Short-Term Memory (ConvLSTM) system for HAR. Initially, individual recognition and posture prediction were applied to precompute skeleton coordinates from the picture and video sequences. Then, the actual skeleton coordinates were exploited with their geometric and kinematic traits to create the new reference traits by the learned ConvLSTM ensemble. Moreover, a categorizer head with a fully connected unit was employed for HAR. But it needs more geometric features and spatiotemporal information to enhance HAR efficiency.

Wu et al. [26] demonstrated a multimodal 2-stream 3D network model for spatiotemporal multimodal training using depth and posture information. Initially, discriminative video representations were constructed to define the spatiotemporal dynamics of action in depth frames by gradually fusing the local movement data. Then, a multimodal 2-stream 3D CNN was applied to train such dynamics. Moreover, the results from each stream were merged for HAR. But its performance was degraded since it did not learn the local spatiotemporal traits of individual activities. Also, the computational cost was high.

Cheng et al. [27] designed an efficient deep ConvNet model for HAR. First, rank pooling was used to extract the spatiotemporal features from the entire RGB-D frame. Then, a twin ConvNet with a cross-modality reward unit was applied to train the cross-modality complementary characteristics and the compensation features from the RGB-D modalities for improving recognition efficiency. But it did not train more comprehensive spatiotemporal traits from various sequences, and it was not effective at fusing the complementary features of multiple modalities.

III. PROPOSED METHODOLOGY

This section describes the GCN model with the JTDGPAHBRD for HAR in detail. Fig. 1 shows an entire pipeline of the study.

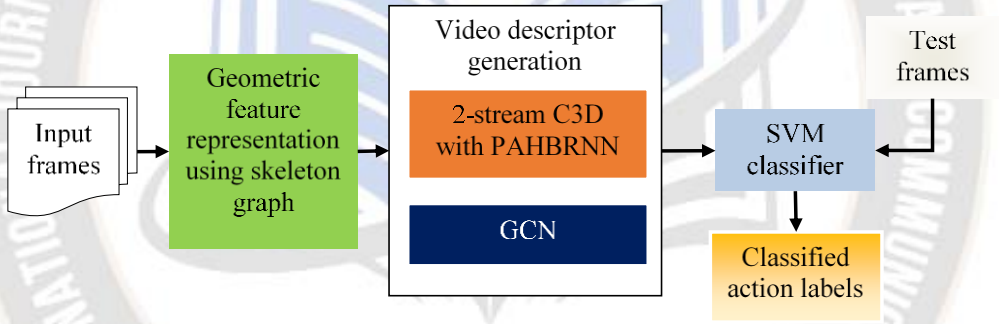

Figure 1. Overall Pipeline of the Study

A. Graph Convolutional Network for Spatial-Temporal Feature Learning

Consider the skeleton information used in the GCN is represented as a spatial-temporal graph $G = (V, E)$ with n skeleton geometries and t frames. So, the skeleton structure's feature map is represented by $X \in \mathbb{R}^{n \times t \times c}$, with c channels defining the joint coordinates.

Normally, in spatial graph convolutions, an adjacency matrix A and an identity matrix I are utilized to delineate the intra-body joint relations that may be split into 3 sets s (about the group of adjacent ensuing from the spatial alignment), where $A + I = \sum_s A_s$. In a specific frame, the graph convolution is defined as follows:

$$Y = \sum_{i=1}^{s} \Lambda_i^{-\frac{1}{2}} A_i \Lambda_i^{-\frac{1}{2}} X W_i \quad (1)$$

In Eq. (1), the degree matrix $\Lambda_s^{ii} = \sum_j (A_s^{ij})$ is the sum of edges linked to every joint node, to regularize A_s, and W_i are the combined weight vectors for all s.

In this study, a dynamic and learnable GCN with a search strategy is applied to create dynamic graphs depending on the node correlations. This mainstreams the temporal dynamics over the respected time-based receptive areas from the temporal graph convolutions, which may be represented as a trainable temporal A. According to this, the correlation matrix S defines the temporal variance for all frames regarding each other. But, because several GCNs are stacked to extract high-level spatial-temporal features, different layers have multi-level semantic data, and the raw integration of S would result in inflexible and fixed temporal configuration to all layers. To avoid this, it is considered necessary to get the geometric traits to various semantic levels. As a result, the convolutional layer is used over

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

the correlation descriptor to train the optimal temporal alignment that best fits the hierarchical GCNs. So, the main contribution is described by

$$R_k = ((\mathrm{S}_k W)^\top \mathrm{J}_k)^\top \quad (2)$$

In Eq. (2), S_k is the corresponding temporal dynamics of the similarity matrix with respect to the temporal indices of the kernel patch k, and J_k is an all-ones vector of dimension c. As a result, the geometric descriptor is dynamically optimized and the temporal descriptor is acquired for each feature channel in the graph convolution by adding (2) to (1), using the Hadamard product:

$$Y = \sum_{i=1}^{s} \Lambda_i^{-\frac{1}{2}} A_i \Lambda_i^{-\frac{1}{2}} (X \odot R) W_i \quad (3)$$

Search Strategy for Dynamic Graph Generation in GCN

Consider a series of $G = \{G_1, \dots, G_T\}$, where all Gs define a skeleton at a spedified interval. The nodes and edges in G define the skeleton joints and their edges, correspondingly. To automatically generate graphs for different layers at different semantic levels, the proposed GCN is integrated with the graph structure search strategy. Initially, the GCN search space constructed with several Gs is defined. After that, a sampling and computation-effective exploration policy is discussed.

GCN search space: A graph search space in the graph structure search strategy defines what and how graph functions an exploring policy might use to construct the GCN. Here, the space assembled with many GCNs is searched to find the best GCN for an adaptive G at various interpretation levels. Types of correlations determined to generate the adaptive G are the following:

1. Topology interpretation relationship: The topology relationship is used to design graph structure according to the current node relations. To define how robust the relationship is between 2 nodes, a standardized Gaussian function is utilized on the graph nodes, and the relationship score acts as a similarity, i.e.,

$$\forall i, j \in V, A_D(i,j) = \frac{e^{\Phi(h(x_i)) \otimes \psi(h(x_j))}}{\sum_{j=1}^{n} e^{\Phi(h(x_i)) \otimes \psi(h(x_j))}} \quad (4)$$

This element is called spatial m. Here, the relationship score $A_D(i,j)$ between node i and j is calculated according to their interpretations $h(x_i)$ and $h(x_j)$. Also, $\otimes$ is the matrix multiplication, $\Phi(\cdot)$ and $\psi(\cdot)$ are 2 estimation parameters, which are applied by the channel-wise convolution filters. According to this, the correlation among nodes is obtained to create the adaptive G.

2. Temporal interpretation relationship: The temporal data of each node is extracted by applying two temporal convolutions before calculating node relationships with Eq. (4). In this manner, the node interfaces among adjacent frames are engaged while computing the node relations. Additionally, a Gaussian function is adopted, as in Eq. (4), to calculate the node relationship. This element is called temporal m, wherein $\Phi(\cdot)$ and $\psi(\cdot)$ are applied by the temporal convolutions.

Using both m, an adaptive G is constructed to learn the spatiotemporal features.

GCN Search Strategy: To reduce the computational complexity of many graphs, the best graph structure must be explored. On the other hand, it is said that various feature layers comprise multiple levels of semantic information, and so a layer-definite strategy is used to create a graph. Therefore, an entire GCN network is searched using this highly computationally efficient search strategy. It finds an optimal structure by estimating the structure distribution. Also, memory efficiency is improved through triggering one function element at all search steps. This exploration policy incorporates a cross-entropy scheme with significance-mixing, where structure variables α is considered as a population and the structure distribution is designed by the Gaussian distribution. After that, this scheme samples a set of structures and using their efficiencies, essential examples are chosen to modify structure distribution. So, the best structure is sampled from the structure distribution.

Initially, the structure distribution is modeled with a Gaussian distribution $\pi \sim \mathcal{N}(\mu, \Sigma)$ and N structure examples $S_{new} = \{\alpha_n^i\}_{i=1}^N$ are sampled as the populations for this scheme. After that, combining S_{new} with the past chosen populations $S_{old} = \{\alpha_o^i\}_{i=1}^N$, an importance-mixing scheme is applied to each population to select structure examples. At last, the freshly chosen examples are utilized to modify the structure distribution π.

During the population selection process, for every population in S_{old} and S_{new}, its probability density in both π_{new} and π_{old} probability density functions (pdf) are compared. So, for the old population α_o^i,

$$\min\left(1, \frac{p(\alpha_o^i; \pi_{new})}{p(\alpha_o^i; \pi_{old})}\right) > r_1 \quad (5)$$

In Eq. (5), r_1 is the threshold randomly selected between 0 and 1, and $p(\cdot; \pi)$ is a pdf with a particular π. Similarly, for fresh example α_n^i from the present distribution,

$$\max\left(0,1, \frac{p(\alpha_o^i; \pi_{old})}{p(\alpha_o^i; \pi_{new})}\right) > r_2 \quad (6)$$

In Eq. (6), r_2 is the other threshold in $[0,1]$. For the modification process, the examples chosen in the past step are utilized to modify mean μ and covariance Σ. Beforehand, the model parameters are modified with the current structure $\alpha = \mu$. After that, the network parameter is predetermined and each chosen example is allocated to the present structure. Using their performances, each chosen example is sorted. According to the efficiency rank, a significance weight λ_i is allocated to the i^{th} example, i.e.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
Article Received: *09 June 2023* ***Revised:*** *28 July 2023* ***Accepted:*** *13 August 2023*

$$\lambda_i = \frac{\log^{(1+N)}/_i}{\sum_{i=1}^{N} \log^{(1+N)}/_i} \quad (7)$$

Accordingly, the example having good efficiency can be provided with a larger weight, which contributes more modification to the distribution. At last, the weighted examples are used to modify the structure distribution, i.e.,

$$\mu_{new} = \sum_{i=1}^{N} \lambda_i \alpha^i \quad (8)$$

$$\Sigma_{new} = \sum_{i=1}^{N} \lambda_i (\alpha^i - \mu)^2 + \epsilon \mathcal{J} \quad (9)$$

In Eq. (9), $\epsilon \mathcal{J}$ is a noise term to better search the graph structure. Because Σ is highly large to determine and modify, it is limited to a diagonal one. Observe that in Eq. (9), the mean of the final iteration is used to modify Σ because the covariance matrix adaption evolution strategy exhibits it is highly effective. The structure of the JTDGPAHBRD-GCN model for video descriptor generation is depicted in Fig. 2. Thus, this GCN using dynamic graphs can capture spatial-temporal features from the skeleton geometries.

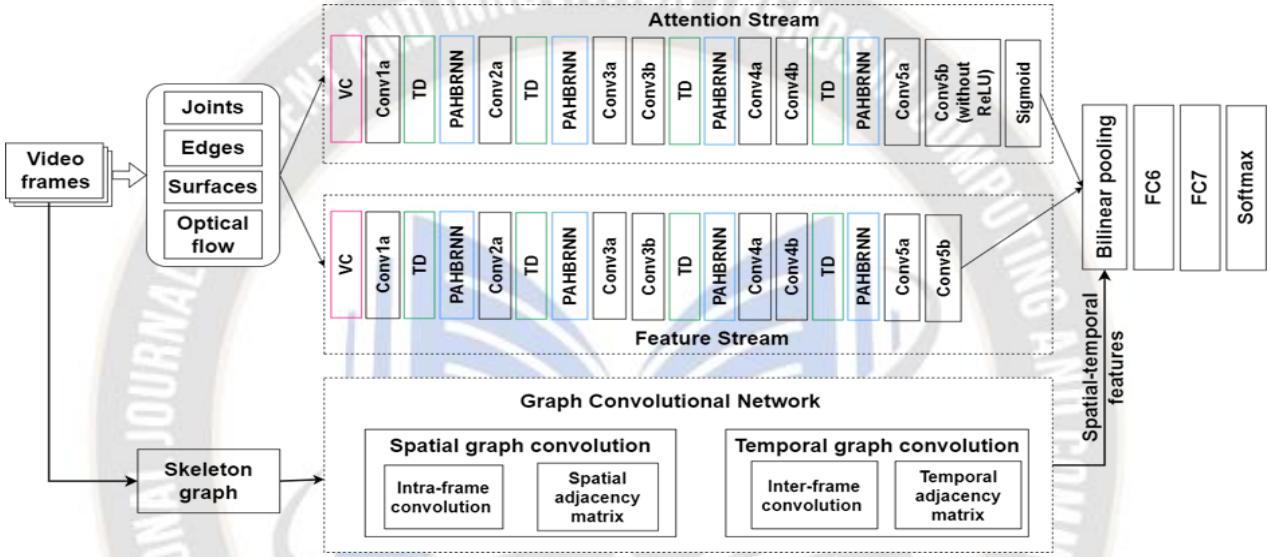

Figure 2. Structure of proposed JTDGPAHBRD-GCN Model Video Descriptor Generation

B. Effective Video Descriptor Generation and Human Action Recognition

After obtaining the complementary high-level spatial-temporal features from the GCN, these features are fused with the body joints, and their trajectory coordinates, which are extracted by the JTDGPAHBRD using a bilinear product. Then, the fused feature vectors are given to the fully connected layer to generate effective video descriptors for particular video sequences. Finally, the generated video descriptors are classified by the SVM classifier into different classes of human actions.

IV. EXPERIMENTAL RESULT

The performance of the JTDGPAHBRD-GCN model is measured in the MATLAB 2017b using the PAD that encompasses 2326 video sequences, each has 15 action classes. All clips are assembled from several web video libraries and involve 50–100 blocks, each of which has 13 body joints annotated. With this dataset, 1861 video sequences are utilized for learning, while 465 video sequences are utilized for testing. Sources include C3D features, coordinates of primitive geometries, trajectory coordinates, and spatial-temporal correlations. To measure the recognition accuracy of JTDGPAHBRD-GCN using these characteristics, several fusion configurations are applied.

The ratio of the number of individual's action classes, which are properly classified is called recognition accuracy.

$$Accuracy = \frac{Number\ of\ recognized\ actions}{Total\ number\ of\ actions\ tested} \times 100\% \quad (10)$$

The example input video frame and its corresponding skeleton image for extracting geometric features and spatial-temporal features are displayed in Fig. 3.

The recognition accuracy results of the JTDGPAHBRD-GCN on the PAD are presented in Table 1.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

(a) (b)

Figure 3. (a) Input frame, and (b) corresponding skeleton image

Table 1. Recognition Accuracy (%) of Sources and JTDGPAHBRD-GCN with Different Alignments on PAD

	Aggregate all the features	JTDGPAHBRD-GCN Ratio Scaling (1×1×1)	JTDGPAHBRD-GCN Coordinate Mapping (1×1×1)	JTDGPAHBRD-GCN Ratio Scaling (3×3×3)	JTDGPAHBRD-GCN Coordinate Mapping (3×3×3)
Geometry features + trajectory coordinates + spatial-temporal features	74.65	-	-	-	-
$fc7$	83.96	-	-	-	-
$fc6$	85.74	-	-	-	-
$conv5b$	82.41	90.11	94.86	89.96	93.08
$conv5a$	73.68	85.35	88.78	84.15	85.44
$conv4b$	65.31	87.19	85.97	88.66	89.23
$conv3b$	54.02	80.54	79.08	80.73	79.67

In Table 1, the first column denotes the accuracy of recognizing human actions using different features such as geometries of the body joints, trajectory coordinates, and spatial-temporal information. It notices that the accuracy of recognizing human actions from the simple aggregation of different features is not satisfactory. Thus, to increase accuracy, each feature from the different layers must be aggregated. $fc7$'s accuracy is slightly lower than $fc6$'s accuracy. It is encouraging since the real C3D-GCN can't alter $fc7$, which is essential to generate an effective video descriptor. Because the geometries and trajectory coordinates of the body joints are used, the results of the PAHBRNN-based pooling at each 3D $conv$ units in JTDGPAHBRD-GCN are examined.

It is observed that when aggregating geometries and trajectory coordinates of the body joints along with the spatial-temporal features in video patterns according to separate parts of the human body (e.g., right leg, right arm, trunk, left leg, and left arm), the JTDGPAHBRD-GCN outperformed the other HAR systems.

Also, JTDGPAHBRD-GCNs from several $conv$ units are combined to determine whether they can balance one another. The outcomes of various configurations applying late merging and the SVM grades on the PAD are shown in Table 2. It compares the accuracy of the JTDPAHBRD-GCN model with the existing models: JTDGPAHBRD [18], TSCN [20], Hyper-GNN [22], and Sybio-GNN [23].

Table 2. Recognition Accuracy (%) of JTDGPAHBRD-GCN by Fusing Different Layers for PAD

Concatenation Layers + GCN	**TSCN [20]**	**Hyper-GNN [22]**	**Sybio-GNN [23]**	**JTDGPAHBRD [18]**	**JTDGPAHBRD-GCN**
$conv5b + fc6$	84.36	86.91	88.25	90.60	93.14
$conv5b + conv4b$	94.10	95.39	97.05	99.70	99.82
$conv5b + conv3b$	83.64	85.26	87.48	90.40	92.51

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

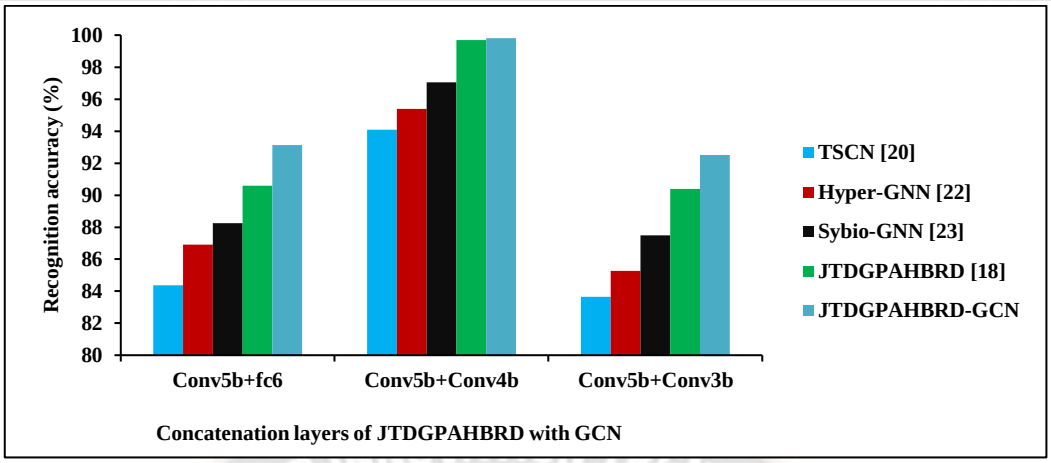

Figure 4. Recognition Accuracy of JTDGPAHBRD-GCN on PAD

Fig. 4 displays that concatenating $conv5b + conv4b$ features in the JTDGPAHBRD with the GCN features has a maximum recognition accuracy compared to the other combinations for feature aggregation. Therefore, it is determined that the JTDGPAHBRD-GCN model can precisely classify human actions in specific video sequences compared to the other existing models. For various HAR models on the PAD, Table 3 provides the performance outcomes of the extracted Geometries+Trajectories+Spatial-Temporal (GTST) features vs. ground-truth GTST features.

The JTDGPAHBRD-GCN achieves a minimum difference between the extracted GTST and ground-truth GTST, as displayed in Fig. 5. From these analyses, it is noticed that the proposed JTDGPAHBRD-GCN model achieved the greatest recognition performance compared to the other HAR models tested by the PAD.

Table 3. Effect of Extracted GTST vs. Ground-truth GTST for Different HAR Models on PAD

Models	Ground-truth	Extracted	Difference
TSCN [20]	0.826	0.801	0.025
Hyper-GNN [22]	0.849	0.833	0.016
Sybio-GNN [23]	0.865	0.851	0.014
JTDGPAHBRD ($conv5b$) [18]	0.893	0.886	0.007
JTDGPAHBRD ($conv5b$)-GCN	0.927	0.922	0.005

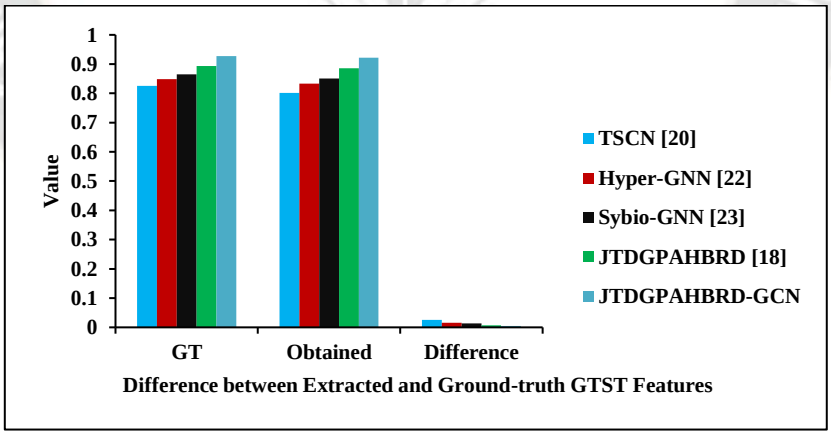

Figure 5. Effect of Extracted GTST vs. Ground-truth GTST for Different HAR Models on PAD

V. CONCLUSION

In this paper, the GCN model was incorporated with the JTDGPAHBRD for achieving spatial-temporal information learning from the skeleton graph. The GCN was used to capture high-level spatial-temporal features between consecutive frames. The search space with multiple dynamic graph structures of the GCN model was created and optimized based on the computation-efficient evolution scheme to learn the temporal dynamics of the skeleton pattern. These newly created spatial-temporal features were aggregated with the geometric features of the body joints and their trajectory coordinates learned by the JTDGPAHBRD for generating video descriptors. Moreover, the obtained video descriptor was classified by the SVM classifier

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

for HAR. Finally, the extensive analysis demonstrated that the JTDGPAHBRD-GCN model on the PAD has a recognition rate of 99.82% by aggregating features of the $conv5b$ and $conv4b$ layers with the GCN features when compared to the other HAR models.

REFERENCES

[1] M. H. Arshad, M. Bilal and A. Gani, "Human activity recognition: review", taxonomy and open challenges. Sensors, vol. 22, no. 17, pp. 1-33, 2022.

[2] M. G. Morshed, T. Sultana, A. Alam and Y. K. Lee, "Human action recognition: a taxonomy-based survey, updates, and opportunities", Sensors, vol. 23, no.4, pp. 1-40, 2023.

[3] R. Singh, A. K. S. Kushwaha and R. Srivastava, "Recent trends in human activity recognition–a comparative study", Cognitive Systems Research, vol. 77, pp. 30-44, 2023.

[4] A. Hussain, T. Hussain, W. Ullah and S. W. Baik, "Vision transformer and deep sequence learning for human activity recognition in surveillance videos", Computational Intelligence and Neuroscience, vol. 1-10, 2022.

[5] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan and A. A. Abbasi, "Human action recognition using fusion of multiview and deep features: an application to video surveillance", Multimedia Tools and Applications, pp. 1-27, 2020.

[6] S. Zhang, Y. Li, S. Zhang, F. Shahabi, S. Xia, Y. Deng and N. Alshurafa, "Deep learning in human activity recognition with wearable sensors: a review on advances", Sensors, vol. 22, no. 4, pp. 1-43, 2022.

[7] Z. Ma, "Human action recognition in smart cultural tourism based on fusion techniques of virtual reality and SOM neural network", Computational Intelligence and Neuroscience, 2021, pp. 1-12.

[8] N. Zhang, T. Qi and Y. Zhao, "Real-time learning and recognition of assembly activities based on virtual reality demonstration", Sensors, vol. 21, no. 18, pp. 1-15, 2021.

[9] Y. Ji, Y. Yang, F. Shen, H. T. Shen and X. Li, "A survey of human action analysis in HRI applications", IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 7, pp. 2114-2128, 2019.

[10] Tiwari, A. K. ., Mishra, P. K. ., & Pandey, S. . (2023). Optimize Energy Efficiency Through Base Station Switching and Resource Allocation For 5g Heterogeneous Networks. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 113–119. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2483

[11] 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.

[12] Mr. Kaustubh Patil. (2013). Optimization of Classified Satellite Images using DWT and Fuzzy Logic. International Journal of New Practices in Management and Engineering, 2(02), 08 - 12. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/15

[13] F. Kulsoom, S. Narejo, Z. Mehmood, H. N. Chaudhry, A. Butt and A. K. Bashir, "A review of machine learning-based human activity recognition for diverse applications", Neural Computing and Applications, pp. 1-36, 2022.

[14] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An and G. Fortino, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: state-of-the-art and research challenges", Information Fusion, vol. 80, pp. 241-265, 2022.

[15] M. Feng and J. Meunier, "Skeleton graph-neural-network-based human action recognition: a survey", Sensors, vol. 22, no. 6, pp. 1-52, 2022.

[16] Anna, G., Jansen, M., Anna, J., Wagner, A., & Fischer, A. Machine Learning Applications for Quality Assurance in Engineering Education. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/109

[17] C. Cao, Y. Zhang, C. Zhang and H. Lu, "Body joint guided 3-D deep convolutional descriptors for action recognition", IEEE Transactions on Cybernetics, vol. 48, no. 3, pp. 1095-1108, 2018.

[18] N. Srilakshmi and N. Radha, "Body joints and trajectory guided 3D deep convolutional descriptors for human activity identification", International Journal of Innovative Technology and Exploring Engineering, vol. 8, pp. 1016-1021, 2019.

[19] N. Srilakshmi and N. Radha, "Deep positional attention-based bidirectional RNN with 3D convolutional video descriptors for human action recognition", In IOP Conference Series: Materials Science and Engineering, IOP Publishing, vol. 1022, no. 1, pp. 1-10, 2021.

[20] Ali Ahmed, Machine Learning in Healthcare: Applications and Challenges , Machine Learning Applications Conference Proceedings, Vol 1 2021.

[21] S. Nagarathinam and R. Narayanan, "Deep positional attention-based hierarchical bidirectional RNN with CNN-based video descriptors for human action recognition", International Journal of Intelligent Engineering & Systems, vol. 15, no. 3, pp. 406-415, 2022.

[22] N.Srilakshmi and N.Radha , "An Enhancement of Deep Positional Attention-Based Human Action Recognition by Using Geometric Positional Features", Indian Journal of Science and Technology,vol. 16,no. 29,pp.2190-2197,2023.

[23] Y. Wan, Z. Yu, Y. Wang and X. Li, "Action recognition based on two-stream convolutional networks with long-short-term spatiotemporal features", IEEE Access, vol. 8, pp. 85284-85293, 2020.

[24] Q. Li, W. Yang, X. Chen, T. Yuan and Y. Wang, "Temporal segment connection network for action recognition", IEEE Access, vol. 8, pp. 179118-179127, 2020.

[25] Z. Ren, Q. Zhang, J. Cheng, F. Hao and X. Gao, "Segment spatial-temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition", Neurocomputing, vol. 433, pp. 142-153, 2021.

[26] X. Hao, J. Li, Y. Guo, T. Jiang and M. Yu, "Hypergraph neural network for skeleton-based action recognition", IEEE Transactions on Image Processing, vol. 30, pp. 2263-2275, 2021.

[27] Singh, M. ., Angurala, D. M. ., & Bala, D. M. . (2020). Bone Tumour detection Using Feature Extraction with Classification by Deep Learning Techniques. Research Journal of Computer Systems and Engineering, 1(1), 23–27. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/21

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10s
DOI: https://doi.org/10.17762/ijritcc.v11i10s.7700
***Article Received:** 09 June 2023 **Revised:** 28 July 2023 **Accepted:** 13 August 2023*

[28] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang and Q. Tian, "Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3316-3333, 2021.

[29] J. Cha, M. Saqlain, D. Kim, S. Lee, S. Lee and S. Baek, "Learning 3D skeletal representation from transformer for action recognition", IEEE Access, vol. 10, pp. 67541-67550, 2022.

[30] S. K. Yadav, K. Tiwari, H. M. Pandey and S. A. Akbar, "Skeleton-based human activity recognition using ConvLSTM and guided feature learning", Soft Computing, vol. 26, no. 2, pp. 877-890, 2022.

[31] H. Wu, X. Ma and Y. Li, "Spatiotemporal multimodal learning with 3D CNNs for video action recognition", IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 1250-1261, 2022.

[32] J. Cheng, Z. Ren, Q. Zhang, X. Gao and F. Hao, "Cross-modality compensation convolutional neural networks for RGB-D action recognition", IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 1498-1509, 2022.