Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16001-9

Human skeleton pose and spatio-temporal feature-based activity recognition using ST-GCN

Mayank Lovanshi[1] · Vivek Tiwari[1,2]

Received: 23 September 2022 / Revised: 3 April 2023 / Accepted: 4 June 2023

Abstract
Skeleton-based Human Activity Recognition has recently sparked a lot of attention because skeleton data has proven resistant to changes in lighting, body sizes, dynamic camera perspectives, and complicated backgrounds. The Spatial-Temporal Graph Convolutional Networks (ST-GCN) model has been exposed to study spatial and temporal dependencies effectively from skeleton data. However, efficient use of 3D skeleton in-depth information remains a significant challenge, specifically for human joint motion patterns and linkages information. This study attempts a promising solution through a custom ST-GCN model and skeleton joints for human activity recognition. Special attention was given to spatial & temporal features, which were further fed to the classification model for better pose estimation. A comparative study is presented for activity recognition using large-scale databases such as NTU-RGB-D, Kinetics-Skeleton, and Florence 3D datasets. The Custom ST-GCN model outperforms (Top-1 accuracy) the state-of-the-art method on NTU-RGB-D, Kinetics-Skeleton & Florence 3D dataset with a higher margin by 0.7%, 1.25%, and 1.92%, respectively. Similarly, with Top-5 accuracy, the Custom ST-GCN model offers results hike by 0.5%, 0.73% & 1.52%, respectively. It shows that the presented graph-based topologies capture the changing aspects of a motion-based skeleton sequence better than some of the other approaches.

Keywords Activity recognition · Pose estimation · ST-GCN · Spatio-temporal feature · Skeleton joints

1 Introduction

Non-intrusive human action identification has gotten a lot of interest from computer vision researchers as it offers numerous emerging applications [1]. It has enough potential to boost applications like video surveillance, health monitoring system, human-computer interactions, the medical field, etc. Several problems, including Several problems, including viewpoint

✉ Vivek Tiwari
viveknitbpl@gmail.com

[1] International Institute of Information Technology (IIIT), Naya Raipur, India

[2] ABV-Indian Institute of Information Technology & Management, Gwalior, India

Published online: 30 June 2023

variation, occlusion, critical body parts, body size distinction of an item, spatial and temporal variations of activities, etc., remain unsolved despite the extensive study done in this field of computer vision [1, 33, 45]. Traditionally, an RGB-based system has been used to track the everyday activities of humans [45].

Moreover, Intelligent technologies have been installed in homes, hospitals, and manufacturing facilities to improve human living and working places in digital surroundings. RGB cameras are subtle to brightness and colour changes, highly congested, noisy, and occluded settings that make a challenge for identifying action through such data [2, 14]. So, modern depth sensor cameras are imposed to rectify the above challenges. RGB depth map images provide geometric information about the pixels in the images by encoding the distance between surfaces of scene objects from a viewpoint. These RGB-depth images have favourable qualities in that they are resistant to changes in light and are scale and rotation-invariant [21, 33, 45].

The development and selection of features in handcrafted algorithms for detecting human activity are driven by prior knowledge and subject-matter expertise. It may not be accurate and limit the system's functionality. Feature selection in a handcrafted approach requires much effort and time, which can be tedious and challenging. Furthermore, it may struggle to recognise complex activities involving subtle variations in human movement patterns. These approaches are limited in adapting to different contexts, as they are designed based on specific features and assumptions [36, 41–44]. In contrast, skeleton-based human activity recognition models can automatically learn features from raw data, making them more flexible and adaptable to different scenarios. They can also handle complex activities more effectively, as deep learning techniques can identify and analyse subtle variations in movement patterns.

Consequently, computer vision offers many methods to handle human body-related action recognition like pose estimation, human identity detection, tracking, correspondence & action recognition [18, 46]. Local feature-based and skeleton-based descriptions of the human body are the two basic approaches [30, 53]. Local feature-based methods see space-time characteristics and utilise bag-of-words models to construct spots at point locations to serve as encoded features. This approach is occlusion resistant but needs much processing power and ignores the spatial relation between the human body parts [2, 13]. While the second methods, i.e., skeleton-based approaches, are more capable of detecting features as they represent a human joint's point and encode the link among distinct joints. The skeleton's joints resist changes in scale and light, invariant to the camera's perspective. Figure 1 depicts the skeleton structure obtained from the OpenPose methods [5]. Ultimately, these technologies are best suited for real-time monitoring systems due to the long frame of computation time.

Graph Neural Networks (GNNs) have recently gotten more attention and are effectively employed in various applications, including image categorisation & action classification [4, 24]. Medical disease diagnosis and prognosis play an essential role in medical therapy & disease detection using such deep-learning algorithms. Deep learning algorithms are now making rapid progress in cardiovascular illness diagnosis, Parkinson's disease monitoring, acute myocardial infarction detection, and falling detection. However, chronic and latent disorders like lumbago and neuralgia remain unquantifiable or challenging to cure. As a result, a computer-vision-based skeleton-based posture estimation system can extract these symptoms for medical considerations. Furthermore, skeleton joints can immediately and accurately monitor the physical states of the human body [46, 52].

Human action recognition and its allied applications have gained widespread use in various domains, from healthcare to e-commerce, smart homes, visual sentiment analysis, and surveillance systems. In healthcare, action recognition can assist health practitioners in identifying and diagnosing motion disorders, such as Parkinson's disease, by analysing patients'

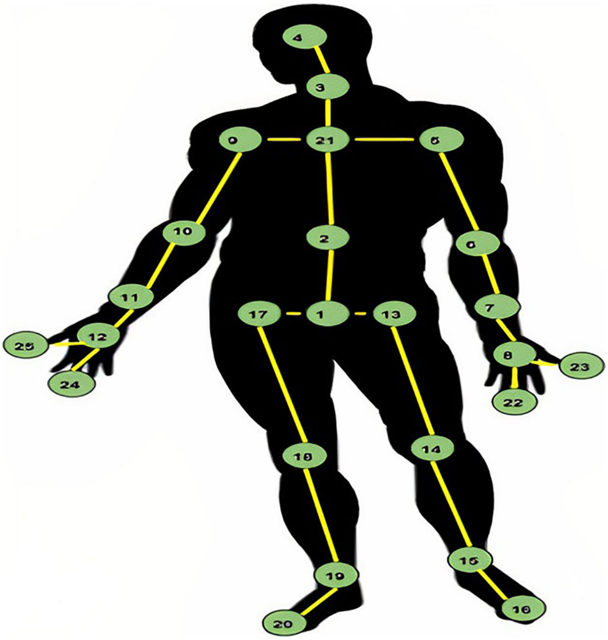


Fig. 1 NTU-RGB-D: human skeleton joint information

movements [39]. In e-commerce, action recognition can enhance the user experience by predicting customers' behaviour and preferences, thereby suggesting personalised recommendations [49]. Smart homes offer various tasks, such as turning on lights or adjusting the temperature based on the occupant's actions [39]. In visual sentiment analysis, action recognition helps interpret an individual's emotional state, enabling businesses to understand the impact of their advertising campaigns [35]. Lastly, abnormal behaviours detection and identifying potential security threats from surveillance videos are established applications [49]. In summary, human action recognition and its allied applications have vast potential to improve various aspects of human life.

This paper projects the graph neural networks to a spatial-temporal graph model termed Spatial-Temporal Graph Convolutional Networks (ST-GCN) to create a standard depiction of skeleton sequences for action recognition. The presented model is based on skeleton graphs, with an individual node representing a human joint. The model includes two forms of edges: spatial edges (which correspond to the intrinsic connection of joints) and temporal edges (link the same joints through successive time steps). On top of it, various layers of ST-GCN are constructed, allowing information to be integrated into both spatial and temporal dimensions [31].

The significant contribution of the present study are:

- A custom spatio-temporal graph-based convolutional network (ST-GCN) has been introduced to recognise the human activity.
- This study attempts a promising solution through a custom ST-GCN model and skeleton joints for human activity recognition where special attention was given to spatial & temporal features.
- Investigate the custom ST-GCN model to extract spatial & temporal features.

- Three human activity datasets, i.e. NTU RGB-D, Kinetics-skeleton & Florence 3D, were employed to validate the model.
- In this paper, the human activity dataset was subjected to normalisation using the Gaussian filter method.
- The proposed model is tested with various performance measures, i.e. Top-1, Top- 5 and Mean loss.

Figure 1 depicts NTU-RGB-D joint information consisting of a set of skeleton sequences and numerous channels. The proposed model attains a significant performance hike compared to the existing method. This approach may useful to solve a medical imbalance of the human body (specifically, pose) through a skeleton dataset since there is always a margin of error between the action recognition of healthy and disabled human body skeleton joints. The deep learning-based pose estimation approach helps us to identify the actions. The following section includes a brief review of skeleton-based human action classification.

2 State-of-the-art

This section discusses recent cutting-edge methods relevant to the proposed work on spatial & temporal feature-based human activity recognition tasks.

[9] offer a framework for view-invariant recognition of human movements with motion and shape temporal dynamics (STD). The proposed RGB Dynamic Images (RGB-DIs) capture the motion content, which is then passed to the optimised InceptionV3 model. The STD stream employs human pose models (HPMs) with view-invariant features to learn long-term view-invariant form dynamics.

A study [29] offers a strongly connected ConvNet with RGB frames and dynamic moving images. A bi-directional long short-term memory (Bi-LSTM) model processes the RGB frames. In the meantime, the lowest layer of the CNN model is trained using a single dynamic motion image. The top layers of the pre-trained model are used to refine the dynamic image stream to extract temporal information from videos.

By examining the impact of the computation of the spatial distribution of gradients (SDGs) on average energy silhouette images (AESIs), the [40] framework seeks to recognise human actions. The AESIs are built to reflect the 3-D pose into a 2-D pose, representing the contour of the action. The SDGs are proposed to compute at various sub-levels. Using the R-transform (RT), the activity's temporal content is determined. The shape of the human body, and temporal evidence, obtained using RT are combined at the recognition stage to create a new, powerful unified feature map model.

[38] integrates the translation and rotation of the human body to recognise human motions. The framework has three key steps: a) the contour of human action is depicted using edge spatial distribution of gradients and directional change of pixel values. b) A transform is used to compute the human action's orientation-based rotational information. c) A descriptor is created by combining the rotational and translational information. This fusion permits the creation of a particular descriptor that can accurately depict the shape and rotational characteristics of the human body during the movement.

The pose-based action features specify human body locations at each frame considered as a feature vector & it was applied throughout the complete activity of human action categorisation. These feature vectors help in machine learning & deep learning-based algorithms that perform regression & classification of action [1]. Human skeleton joints can be extracted manually as well as automatically. The manual method employs skeleton joint annotations, while automatic processes use joint position estimate approaches from a human depth image [1]. RGB-Depth images have been utilised to automatically develop a method for calculating the human skeleton. Many pre-trained libraries in python automatically detect skeleton sequences from depth images like OpenPose, Mediapipe, WrenchAI, etc [5, 7]. Figure 2 depicts a skeleton using the OpenPose library. The OpenPose is a real-time multi-person key-point recognition framework with skeleton information to estimate the body, face, hand, and foot [1].

Four types of pose-based features play a vital role, i.e., movement feature, location feature, raw joint position data & multi-modal feature [1]. The extracted pose description should be presented in the fixed-size feature vector before applying these features to the classifier. Traditionally, extracted features were represented in a code book and used as a visual vocabulary to translate feature vectors depending on their delivery of the learned vocabulary. In action recognition, these feature vectors indicate a subgroup of the pose sequence (key poses), which may be used to recognise an action rather than the entire sequence. This procedure correlates identified postures with a pose sequence in the dictionary and generates a histogram of the key pose sequence. The lack of temporal information in these bag-of-visual-words approaches is a problem [1]. To improve such a problem, numerous methods have been suggested.

[1] suggested using a fisher vector to encode the skeleton features and tries to recognise rapidly and handle missing temporal information problems. Firstly, the skeleton joints are plotted into a three-D space, including their location, velocity, acceleration, and label. Secondly, A Fisher Vector(FV) encoding is finalised to whole sets of skeleton sequences before they are proposed to an additional space as a set. Fisher vector encoding works on Gaussian mixture models(GMM) to develop a visual dictionary. The FV is calculated by identifying the GMM distribution parameters covering the distribution of the supplied descriptors. Fisher vector is a sum of normalised gradient statistics computed for each descriptor. So, the Fisher

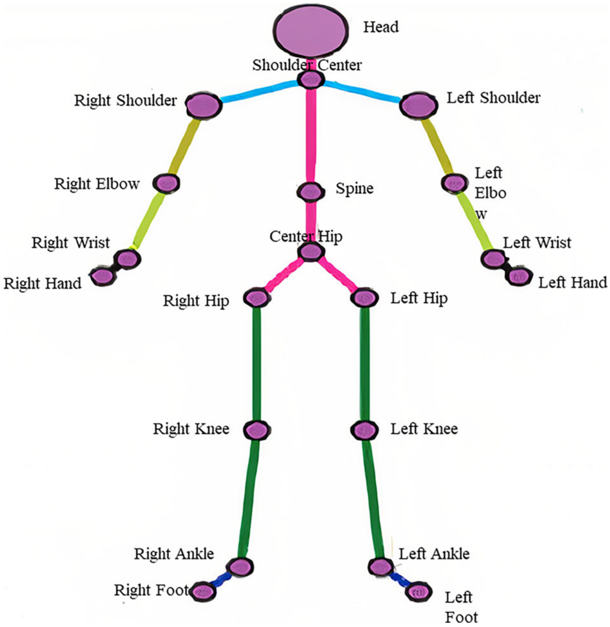


Fig. 2 Sample skeleton images using OpenPose

vector is defined as an Eq. (1) & it is dependent upon the score function. The score function is the partial derivative of the log-likelihood function, i.e., shown in the Eq. (2).

$$\mathrm{FV} = \sum_{t=1}^{T} \mathrm{L}_{\lambda} \nabla_{\lambda} \log u_{\lambda}(x_t) \tag{1}$$

Where,

$$\text{Score function} = \nabla_{\lambda} \log u_{\lambda}(X) \tag{2}$$

Let,

X = Set of D-dimensional local descriptors extracted from an image.
L_{λ} = Chloskey decomposition
u_{λ} = Probability density function
∇_{λ} = Gradient of descriptor

A framework [2] based on Gaussian kernel correlation proposes a single depth sensor to estimate human posture. The kinematic skeleton was included in the Gaussian kernels, and a few multivariate Gaussian kernels expressed the tree-based framework rotated using quaternions. By incorporating regularisation terms, they created a capable and resilient sequence-based pose-tracking system (visibility, continuity and self-intersection).

The study [45] proposed a method for human behaviour analysis using numerous cameras. The system combined each individual's body motion, proximity to others, and sound to create interaction paths. The literature employed kernel-state space models and developed paired kernels through specific proportions to represent the temporal sequence. In this literature, the spatial feature is missing.

In [33], literature discussed an ensemble of vigorous time warping, Fourier temporal pyramid representation, and linear SVM to classify data. This representation outperforms many existing skeleton representations on NTU-RGB-D datasets. Due to the small number of datasets employed by these methodologies, it does not provide a full review of the proposed framework.

[37] presents a method to recognise human activities in 3D using data from depth sensors. The proposed approach creates a movement polygon that captures the spatio-temporal features of the activity using data from the posture skeleton. The researchers suggest a new classification technique that categorises the movement polygon using support vector machines (SVMs).

A Recurrent Neural Networks (RNN) model is proposed [11] for skeleton-based action recognition because RNN can simulate action classification on a skeleton dataset of temporal sequences. RNNs were used in temporal graphs to study action recognition models. In [27], literature simultaneously studies the unknown sources of action-related data within the input data across both Spatial-temporal domains. The present study doesn't discuss the detailed hyperparameter tunning used for experiments.

A framework discussed [6] is based on the Extreme Learning Method (ELM), a machine learning feed-forward technique used for classification and regression. ELM has been worked with a single layer that is rapid, accurate & effectively deployed in various areas with its short training period; it has attained acceptable accuracy on large-size datasets for video classification. ELM is faster, more efficient, and cost-effective than the training time of many deep convolution networks. ELM is best suited for the temporal sequence but not spatial feature extraction. This paper employed ELM approaches, which are challenging to cope with when extracting spatial features.

In [28], the deep learning approach was applied to recognise human action using a past aware LSTM (P-LSTM) algorithm. P-LSTM are more effective since only a portion of the body's motion is needed; this technique does not recall the complete motion. Input, forget, and each part uses modulation gates, but all body parts share the output gate. A method introduced by the [15] model is based on the video frame by averaging the prediction from the RGB frame and a stack of ten externally calculated optical flow frame methods. The two-layer ImageNet convolution techniques processed these inputs. The discussed method only uses the RGB frame that causes noise in the data.

[26] uses a model for recognising human actions based on skeleton joints that are weakly aligned in three dimensions. Each frame's feature vector for the human body includes the joint coordinates and temporal features considered temporal derivatives. These models aid in reducing noise and enabling joint alignment in humans. In this [16], joint information is extracted from the frame using a pre-trained CNN model and a temporal pooling approach. This study collects the CNN feature of three clips with the same timestamp as a single feature vector. The parallel learning of these features uses the Multi-Task Learning Network (MTLN) [16], which is used for motion prediction. The discussed approach is based on the RGB frame, which has challenges like low light, background clutter etc.

A new method was introduced [10] for recognising human actions from 3D video data author argues that recent techniques fail in capturing crucial spatio-temporal information and suggest a novel strategy that uses attention mechanisms to concentrate on pertinent areas of the input data. The proposed approach uses a hybrid 2D-3D CNN architecture with attention mechanisms, focusing on different input data regions at different times. The authors suggest a brand-new loss function considering the attention mechanism's efficiency and classification accuracy.

A convolution encoder-decoder architecture is used in the temporal convolution network [17] to segment temporal motion in the video. The model collects filtered features to predict. [8] described a technique for dividing an action sequence into motion units that combines pose-based and segment-based techniques. A framework used for shape analysis that represents and compares shapes in a Riemannian manifold [8] to evaluate the shape of the human pose and the shape of its motion. The accuracy of this technique can be affected by noise in the data, such as occlusion, lighting changes or errors in pose estimation. These factors can introduce errors in the analysis & affect the reliability of the result.

A low-cost, interactive framework for full-body rehabilitation based on 3D immersive serious games is presented in [3]. Using a Kinect sensor, the technology tracks the patient's movements and converts them into control signals for an avatar in a virtual environment. The virtual environment provides a range of activities that simulate real-world tasks relevant to the patient's rehabilitation needs. The system also offers real-time feedback to patients based on their performance. The study's results showed significant improvements in the patient's range of motion, balance, and gait speed. This article used less sample size, limiting the result's generalizability.

A 3D skeleton joint is depicted [34] as a Euclidean group included within the curved manifold Lie group [34]. The lie algebra serves as the foundation for the action curves. A combination of the Fourier pyramid, linear SVM, and dynamic temporal wrapping is used for classification. [22] employs a graph-based method to teach a data graph with k-connected components for the clustering, which is useful for extracting posture using a graph-based approach. The intrinsic relationship between joint configurations and action classes is identified [50] using a discriminative multi-instance multitask learning (MIMTL) framework. The performance of the discussed framework depends on the effectiveness of the feature extraction process, which may require domain-specific knowledge & expertise. Poor feature extraction may result in the suboptimal performance of the model.

An enhanced skeleton-based graph convolution network (GCN) for human action & interaction was developed, [47] which is based on the extended skeleton graph topology and partitioning strategy to extract a large portion of the non-adjacent joint relational information in the model for robust discriminant features. By utilising ST-GCN, the expanded skeleton graph and partitioning approach are implemented.

Existing approaches for detecting human activity employ a Multi-class classifier, which assigns a 1-of-N class to each action classification. In light of the literature discussion, most studies utilise both spatial and temporal features of human action. It seems to perform better if trying to accommodate the combined effect of both features in the calculation. In this view, the Spatial-Temporal Graph Convolution Networks (ST-GCN) model has been proposed to effectively study spatial and temporal dependencies. The proposed approach is based on the skeleton, which offers better recognition. The skeleton approach has several advantages over dynamic motion image [9], average energy silhouette image (AESI) [40], and gait energy image (GEI) [38] in the context of human activity recognition, such as robustness, generalizability, efficiency, interpretability etc.

The proposed method offers an architecture that captures the spatio-temporal information from data. The approach seems particularly effective in handling issues such as low light, occlusion, background clutter, and errors in pose estimation that can arise in RGB image or video data. Furthermore, using a Gaussian filter to normalise the data is a technique in the proposed model and can help remove noise from the data. The graph convolution network used by the proposed model is an interesting choice. When working with pose estimation data, Graph Convolutional Networks (GCNs), a form of a neural network, can work on graph-structured data. By using GCNs, the proposed method effectively recognise human activities, as these networks can better capture the spatio-temporal relationship.

3 Proposed methodology: Skeleton pose and spatio-temporal feature

A skeleton-based method compound with two feature channels has been presented in this study to describe human activity. The first channel covers the spatial aspect, while the second defines the time aspect of the action. Each activity sequence contains two feature channels: f_s represents a spatial variation of each frame sequence & f_t represents a temporal variation of each sequence of action class. To improve the noise ratio & to smooth the skeleton image data, the Gaussian filter method has been used to grind on skeleton data, as shown in the Eq. (3).

$$G_\sigma = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2+z^2)}{2\sigma^2}} \tag{3}$$

Where,

σ = standard deviation
x = joints value of skeleton on the x-axis
y = joints value of skeleton on the y-axis
z = joints value of skeleton on the z-axis

A raw NTU-RGB-D skeleton image is depicted in Fig. 3, along with the pixel intensity, which scale from 0 to 255. While Fig. 4 shows a pixel-intense image of a smooth skeleton that has undergone Gaussian filtering, which is used to smooth out & reduce noise in an image. The raw skeleton image has 72000 pixels at 255-pixel intensity. In contrast, the smooth

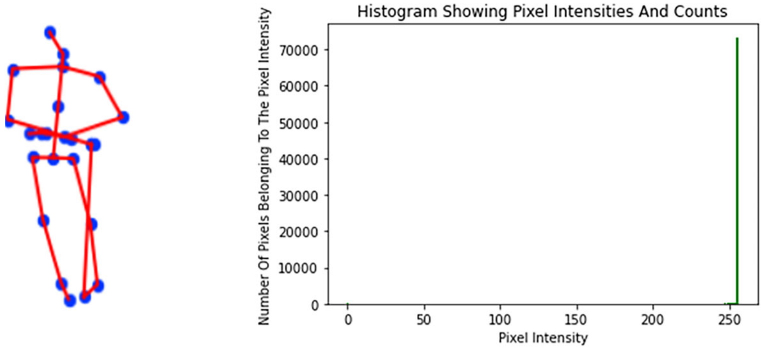


Fig. 3 Pixel intensity histogram: Raw skeleton image

skeleton image with a Gaussian filter offers a 110000-pixel count at a 255-pixel intensity. Gaussian-filtered skeleton images provide less noisy smooth images that improve the input data in a smoothing manner because high pixel intensity counts result in less noise in the image [32]. Therefore, Gaussian-filtered skeleton image is used as input for an experiment because of their smoothness & less noise.

3.1 Spatial feature channel

Spatial features are to be calculated by determining the spatial deviation of each frame of the activity sequence. In this view, a matrix of point position has been used to measure the spatial deviation. The matrix of point position through $Edistance$ (Euclidian distance) & $Cdistance$ (Cosine distance) between joints is shown in Eqs. (4) and (6), respectively. Figure 5 depicts the extracted spatial feature of the sample skeleton image. The complete procedure

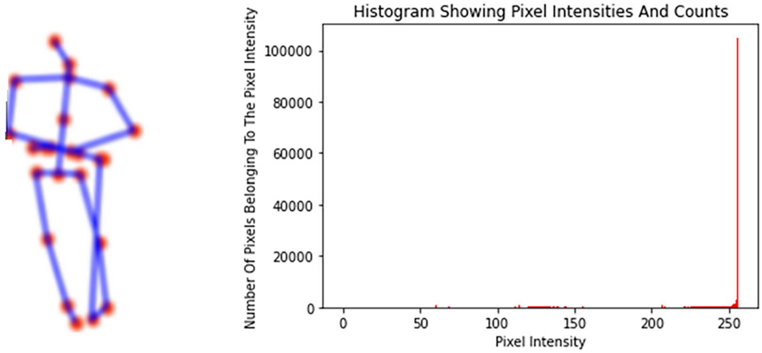
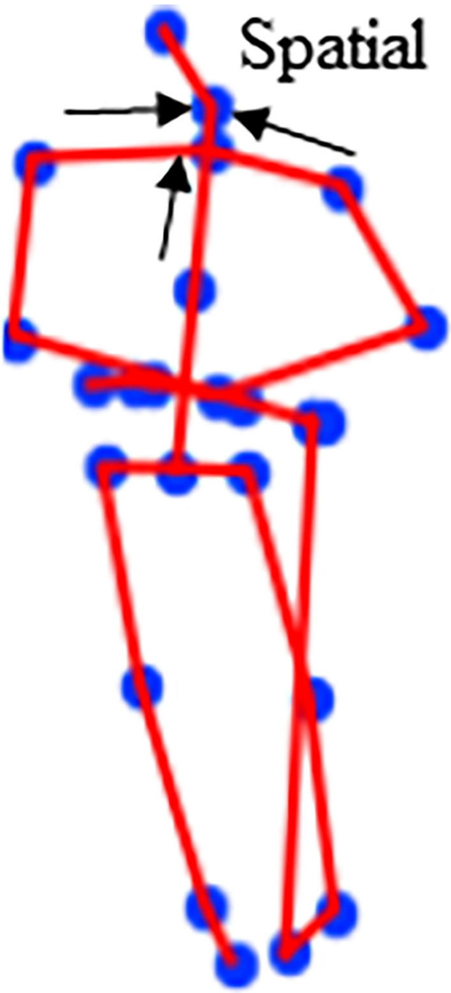


Fig. 4 Pixel intensity histogram: Smooth skeleton image with Gaussian filter

Fig. 5 Spatial feature of skeleton image



is represented in Algorithm 1 (Spatial Feature Generation).

$$E_{distance} = \begin{bmatrix} e_{11} & e_{12} & e_{13} & e_{24} & \dots & e_{1N} \\ e_{21} & e_{22} & e_{23} & e_{24} & \dots & e_{2N} \\ e_{31} & e_{32} & e_{33} & e_{34} & \dots & e_{3N} \\ e_{41} & e_{42} & e_{43} & e_{44} & \dots & e_{4N} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ e_{N1} & e_{N2} & e_{N3} & e_{N4} & \dots & e_{NN} \end{bmatrix} \tag{4}$$

where,

N = Total number of joints in each skeleton image
e_{ij} = Euclidian distance between two joints J_i and J_k

In Eq. (5), Euclidian distance between two places $X = (x_1, x_2, x_3 ... x_n)$ and $Y = (y_1, y_2, y_3 .... y_n)$ €R:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (Y_i - X_i)^2} \tag{5}$$

The matrix of point position using cosine distance is shown in Eq. (6).

$$C_{distance} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{24} & \dots & c_{1N} \\ c_{21} & c_{22} & c_{23} & c_{24} & \dots & c_{2N} \\ c_{31} & c_{32} & c_{33} & c_{34} & \dots & c_{3N} \\ c_{41} & c_{42} & c_{43} & c_{44} & \dots & c_{4N} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ c_{N1} & c_{N2} & c_{N3} & c_{N4} & \dots & c_{NN} \end{bmatrix} \tag{6}$$

Where,

N = number of joints in each skeleton image
C_{ij} = cosine distance between two joints J_i and J_k

Cosine distance between two vectors x, y is shown in Eq. (7):

$$CosineDist(x, y) = 1 - CosineSim(x, y) \tag{7}$$

Let,

$$cosineSim(x, y) = (xy)/(|x||y|) \tag{8}$$

$CosineSim(x, y)$ = cosine similarity between x, y in Eq. (8).
$x.y$ = standard dot product of two vectors x & y.

Equation (9) depicts the spatial features channel calculated by concatenating two $E_{distance}$ and $C_{distance}$.

$$f_s = Concat(E_{distance}, C_{distance}) \tag{9}$$

The total space taken by the spatial variation features channel for the frame is equal to $2 \times N^2$.

3.2 Temporal feature channel

Temporal feature is associated with times and may keep on changing. So, estimating the temporal change of each frame of the activity sequence seems promising. It is measured by computing the change in coordinates between each joint J_i and the maximum and minimum values of the same joint over the whole sequence. A joints J_i with the 3D coordinates (J_{ix},J_{iy},J_{iz}), We calculate J_i,max through Eq. (10) and J_i,min through Eq. (11) as follows:

$$J_i, max = (max(J_{ix(t)}) - J_{ix}) + (max(J_{iy(t)}) - J_{iy}) + (max(J_{iz(t)}) - J_{iz})/3 \tag{10}$$

$$J_i, min = (J_{ix} - min(J_{ix(t)}) + (J_{iy} - min(iy(t)) + (J_{iz} - min(J_{iz(t)})/3 \tag{11}$$

Where,

[$max(J_{ix(t)})$,$max(iy(t))$,$max(J_{iz(t)})$] = maximum value of J_i coordinates of all sequences of the skeleton.

```
Algorithm 1: Spatial Feature Generation
Input:
1. $S$ : Skeleton image
2. $J_i$ : Joints of skeleton image such that
   $J_i \in S$        where $i = 1 ......N$
Output:
1. $F_s$ : Spatial Feature
Function:
1. np.linalg.norm(x,y,z): It calculates Euclidean distance.
2. spatial.distance.cosine(x,y,z): It calculates cosine distance.
3. concat($E_{distance}$, $C_{distance}$): It is used for concatenate $E_{distance}$ & $C_{distance}$
Procedure:
1. For $i$ in range 1 to $max(J_i)$:
      For $j$ in range 1 to $max(J_j)$:
      $e_{ij} \leftarrow$ np.linalg.norm($J_{ij}$)
      $E_{distance} \leftarrow e_{ij}$        $\forall\, e_{ij} \in E_{distance}$
      $C_{ij} \leftarrow$ spatial.distance.cosine($J_{ij}$)
      $C_{distance} \leftarrow c_{ij}$        $\forall\, c_{ij} \in C_{distance}$
      $F_s \leftarrow$ concat($E_{distance}$ , $C_{distance}$)
   End
End
2. Return $F_s$
```

$[min(J_{ix(t)}),min(iy(t)),min(J_{iz(t)})]$ = minimum value of J_i coordinates of all sequences of the skeleton.

Finally, f_t is calculated by adding the max and min values of all joint differences & Eq. (12) represents the temporal feature. Figure 6 depicts the extracted temporal feature of the sample skeleton image. The complete procedure is represented in the Algorithm 2 (Temporal Feature Extraction).

$$f_t = (J_1, min....J_N, min,\ J_1, max.....J_N, max) \tag{12}$$

The total space of temporal variation features is 2 × N.

3.3 ST-GCN model

This study proposed to employ an ST-GCN (Spatial-Temporal Graph Convolution Networks) that uses skeleton data (OpenPose) to recognise human activities. Once trained, it can detect

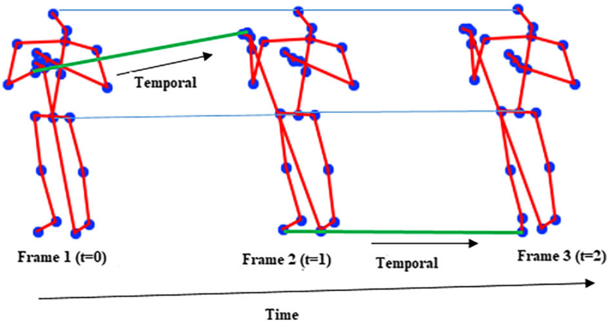


Fig. 6 Temporal feature extraction from skeleton image

Algorithm 2: Temporal Feature Extraction

```
Input:
1. $S$ : Skeleton image
2. $J_{ix}$, $J_{iy}$, $J_{iz}$ = Joints of skeleton image |
$J_{ix}$, $J_{iy}$, $J_{iz} \in S$          where i = 1 ......N & x,y,z is axis of the image
Output:
1. $F_t$ : Temporal Feature
Function:
3. concat(): It is used for concatenating two joints' value
Procedure:
1. For $i$ in range of 1 to $N$:
    $J_{i,max}$ =[(max($J_{ix(t)}$) - $J_{ix}$ + (max($iy(t)$) - $J_{iy}$) + (max($J_{iz(t)}$) - $J_{iz}$)] / 3
End
Return ($J_{i,max}$)
2. For $i$ in range of 1 to $N$:
    $J_{i,min}$=[($J_{ix}$- min($J_{ix(t)}$)) + ($J_{iy}$- min($iy(t)$)) + ($J_{iz}$- min($J_{iz(t)}$))] / 3
End
Return ($J_{i,min}$)
3. $F_t \leftarrow$ concat[($J_{i,max}$ .... $J_{N,max}$) , ($J_{i,min}$ .... $J_{N,min}$)]
Return $F_t$
```

numerous activities with state-of-the-art underlying datasets, like the NTU-RGB-D and [28], Kinetics-skeleton[15] & Florence 3D dataset[19]. ST-GCN uses the OpenPose Algorithm to acquire skeleton-based data from video & images. The data is usually in a series of frames, each with its sequence of joint coordinates. We establish a spatial-temporal network using the 2D and 3D coordinates of the body joint. The joints are graph nodes and intrinsic links in human body structures as graph edges. As a result, the joint coordinate vectors on the graph nodes are the ST-GCN input, and it might be similar to image-based CNNs, which take pixel vectors from a 2D picture grid as input. The input data will be processed through many levels of spatial-temporal graph convolution processes, resulting in higher-level feature mappings on the graph. The standard SoftMax classifier is used to classify it into the appropriate action category. The backpropagation has been employed to train the whole model [51].

The graph can be considered $G(V, E)$, Where V is vertices & E is an edge. It doesn't explicitly consider any features with static structure. At the same time, GNN (Graph neural network) can be treated as G (V, E, f_t, f_s), where f_t & f_s is a static feature with a static structure. However, while dealing with a graph with a static structure and time-varying properties, spatial-temporal GNN is promised to be used. So Spatial-temporal graph can be considered $G(V, E, f_t(t), f_s(t))$ where $f_t(t)$ & $f_s(t)$ is a Spatio-temporal feature with static structure.

This paper employs a deep-learning-based framework (model), i.e., a Custom Spatio-temporal graph convolution network (ST-GCN), to classify action categories. Figure 7 depicts the architecture of the model [51]. The proposed model is designed with eleven layers of the Spatio-temporal convolution unit, starting with three channels, then 64 channels in the next four layers, 128 channels in the following four layers, and 256 channels in the final three layers. The number of input channels depends on the modality of the data, such as RGB images, depth maps, or skeleton data. So, in ST-GCN, the input data consists of 3D joint positions, and each joint position can be represented as a 3D coordinate, resulting in a 3-channel input. The channel count of the model is typically increased in the deeper layers of the network to capture more complex features. The First ST-GCN unit is attached to the batch normalisation (BN) layer, while the last ST-GCN unit is connected with the global average pooling (AP) & softmax layer. It is denoted as BN→11 ST-GCN→AP→SoftMax,

Algorithm 3: Proposed Method: Activity Recognition

```
Input:
1. $R$ : RGB-D image | $f_1$, $f_2$ ... $f_N \in R$        where  $f_1$,$f_2$ is frame of the image
Output:
1. $A$ : Action Category
Function:
1. ST-GCN(): Spatial-Temporal Graph convolution network
2. $F_s$(): Spatial feature using spatial feature extraction algorithm
3. $F_t$(): Temporal feature using temporal feature extraction algorithm
4. concat(): It is used to concatenate spatial & temporal feature
Procedure:
1. For $i$ in range of 1 to $N$:
    skeleton image: $S_i \leftarrow f_i$
End
Return $S_i$
2. For $i$ in range of 1 to $N$:
    $J_{ix}$,$J_{iy}$,$J_{iz} \leftarrow S_i$            where $J_{ix}$,$J_{iy}$,$J_{iz}$ = coordinate of joint on corresponding axis
End
Return $J_{ix}$,$J_{iy}$,$J_{iz}$
3. $F_s \leftarrow J_i$......$J_N$ using a spatial feature algorithm.
4. $F_t \leftarrow J_i$......$J_N$ using a temporal feature algorithm.
5. Spatio-temporal feature : $F \leftarrow$ concat($F_s$,$F_t$)
6. $A \leftarrow$ ST-GCN(F)
Return $A$
```

where SoftMax signifies the softmax activation function layer. The BN layer normalises the feature to a layer for every mini-batch. AP calculates the average value of the feature map & uses it to create a down-sampled (pooled) feature map.

However, every ST-GCN unit was designed with the graph convolution network(GCN) & temporal convolution network (TCN) blocks. The GCN block incorporates a convolution layer only that convolves features & transfers them to the TCN network. In contrast, TCN blocks consist of batch normalisation (BN), ReLu activation layer, convolution layer & dropout layer. It is denoted as BN→ReLu→Conv → BN → Dropout, where dropout is used to avoid over-fitting by reducing the number of layers. TCN is a framework that offers a causal convolution layer and dilations to use its temporality and broad receptive fields to be adaptable for sequential data. A causal convolutional layer applies the kernel only to the past values of the input, ensuring that the output at each time step is causally related to the past information. This means that a causal convolutional layer can be used to model time-series data [51]. In custom ST-GCN, the causal convolution layer extracts spatiotemporal features from graphs over time and enables to capture of the temporal dependencies. This helps the model learn patterns and relationships in the evolving data, improving its ability to make accurate predictions.

A custom ST-GCN model is introduced in this study, consisting of two GCN in place of a single GCN in the ST-GCN unit. Two GCN layers help improve learning and results because a doubly convoluted feature map provides a better input for the TCN network. GCNs capture human activity's spatial and temporal dynamics and the interactions between various body components. The proposed model employed a Graph Attention Mechanism (GAM) to deal with the various temporal lengths of human activities in GCNs. Different temporal action lengths can be effectively managed when GCN is coupled to GAM. Based on the individual activity, GAM assists in learning the attention weights for various nodes and edges. Hence, human activity is classified using the learned node, edge, and attention weights [48].

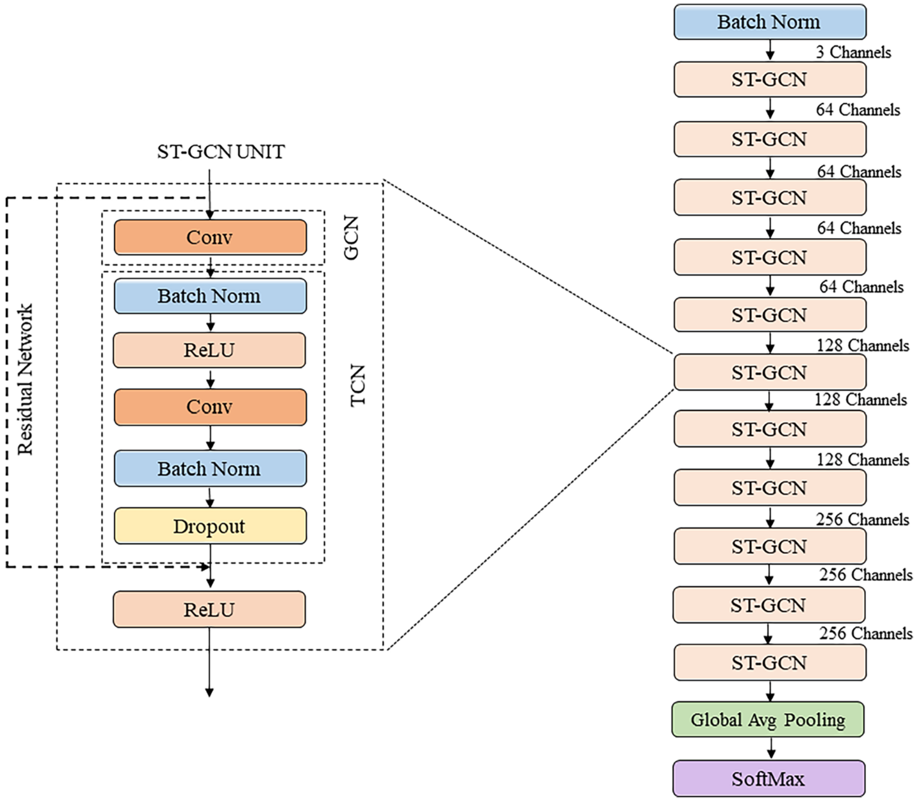


Fig. 7 Custom ST-GCN: Spatio-temporal graph convolution network architecture

A proposed framework is depicted in Fig. 8. Table 1 projects the model description with the hyper-parameter of the custom ST-GCN network. Where SGD, or stochastic gradient descent, is an iterative method for improving an objective function's smoothness properties. SGD addressed the issue of Gradient Descent by updating parameters using just one record. The pseudo-code of the proposed method is represented in the Algorithm 3, which shows the process flow of the custom ST-GCN model. This proposed methodology identifies action categories from the skeleton image using the Spatio-temporal feature (Algorithm 1 & 2).

ST-GCN is substantially quicker than traditional 3D convolution for action detection since it can estimate actions based on skeleton input. Graph convolution also considers the link between joint skeleton points, resulting in greater accuracy than applying 2D convolution to the skeleton information alone. The skeleton-based dataset offers spatio-temporal information, which gives benefits during model training but comes with some challenges. The major issues are related to data pre-processing without impacting spatio-temporal information, keeping the model less complex, deciding channel, interpretability of spatio-temporal features, hyperparameter tuning etc. The Custom ST-GCN offers such an architecture that is able to deal with such issues.

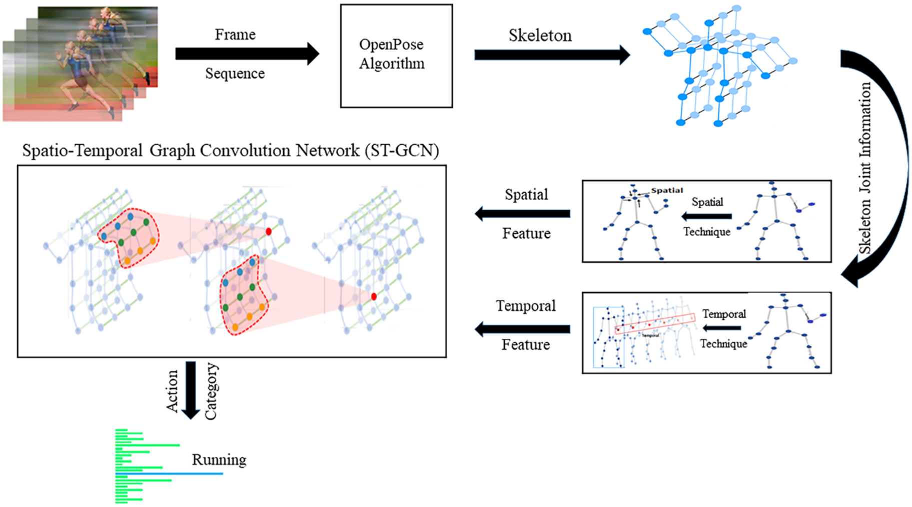


Fig. 8 Proposed Framework: Spatio-Temporal features and custom ST-GCN model

4 Dataset explored

The underlying datasets play a vital role in machine learning, becoming more crucial for skeleton-based activity recognition. There are limited datasets available in this domain. In this view, three state-of-the-art datasets have been used for experiment and validation. Table 2 depicts the dataset's descriptions. The following section is worth discussing the dataset.

4.1 NTU-RGB-D

The NTU-RGB-D 60 dataset is one of this domain's most referenced datasets for action recognition. There are 60 action classes termed A1 to A60 (clips). Where 40 classes are

Table 1 Model Description

SN	Hyperparameter	Values
1	Epoch	80
2	# of GPU	2
3	Batch size	256
4	Test batch size	64
5	Channel	3
6	# of class	60
7	Dropout	0.5
8	Learning rate	0.01
9	Optimiser	SGD
10	Weight decay	0.0001
11	Layer	11 Convolutional layer
12	Model(in convolution)	RESNET

Table 2 Dataset description of skeleton dataset

SN.	Parameter	NTU-RGB-D	Kinectics-400	Florence- 3d
1	# of clips	56880	306245	21864
2	Size of dataset	5.8 GB	42.2 GB	2.1 GB
3	# of Activity/ classes	60	400	215
4	# of joints	25	20	15
5	# of body information	10	-	10

daily actions(drinking water, picking up, writing, jumping up etc.), 9 are medical actions (sneezing, neck pain, yawning etc.) & the remaining is two-person interaction activities (kicking, pushing, shaking hands). Many terminologies are used by other research for these datasets, such as NTU-RGB-D, NTU RGB-D, NTU RGBD etc. Table 2 depicts a brief description of the dataset. NTU-RGB-D offers two benchmarks, i.e., cross-subject(x-sub) & cross view (x-view), which have different classifications between test & train split of total clips. Forty people acted in the NTU-RGB-D dataset, dividing 20 people for the training dataset and 20 for the testing sets. NTU-RGB-D collected information via three cameras; the first camera was for the testing data collection, and the second & third cameras were for the training data collection. The cross-subject has 40360 clips for training & 16560 clips for testing evaluation, while the cross-view has 37920 for training & 18960 clips for testing evaluation. A few sample images of cross-subject are shown in Fig. 9, with OpenPose 2D skeleton. The dataset is available on https://rose1.ntu.edu.sg/dataset/ [28].

4.2 Kinetics-Skeleton

The dataset contains 400 human activity classes containing 765 clips per class, each around ten seconds long. The dataset is emphasised human activities. The current version comprises

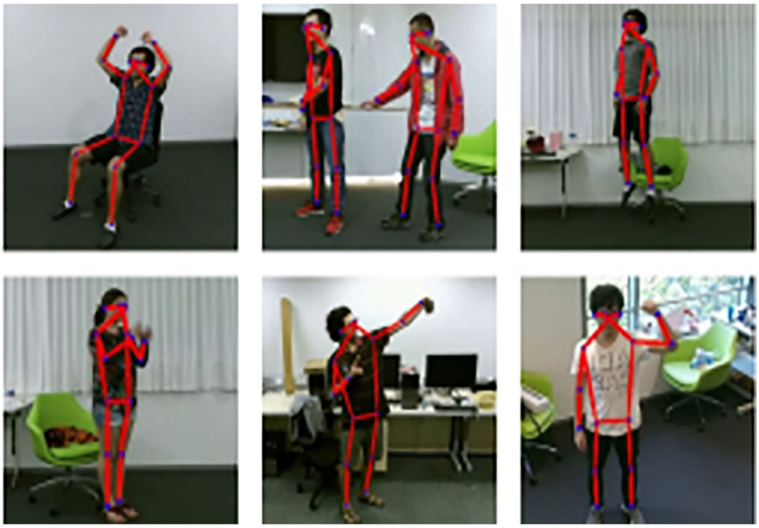


Fig. 9 The extracted OpenPose 2D skeleton from the NTU-RGB-D dataset

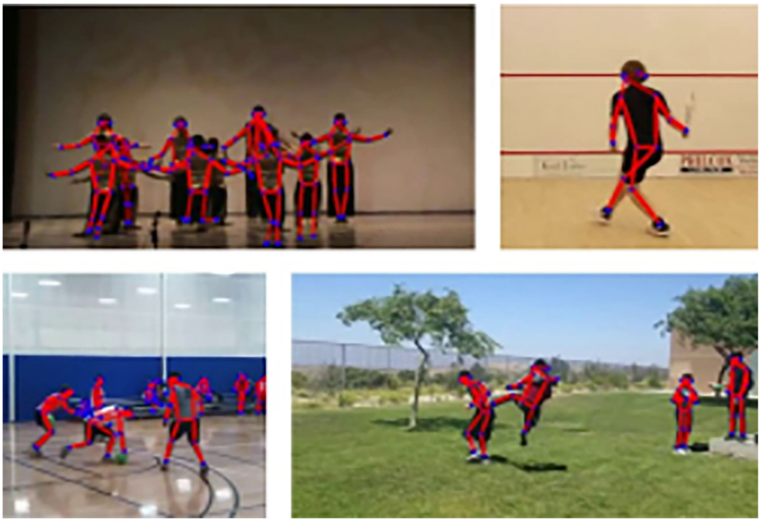


Fig. 10 The extracted OpenPose 2D skeletons: a sample from the Kinetics-Skeleton dataset

306,245 clips that are divided into three categories: training (500 clips per class), validation (100 clips per class), and testing (165 clips per class). The clips are from YouTube videos, and their resolution and frame rates continuously vary. Human actions (singular) are included in the action classes, e.g., sketching, drinking, laughing, beating, hugging, kissing, shaking hands, moving the yard, and cleaning dishes. Some actions, such as swimming, are fine-grained and require temporal thinking to discriminate between them. Other actions, such as playing different wind instruments, necessitate a greater focus on identifying the thing. Figure 10 shows a few sample images with OpenPose 2D skeleton. The URL of the YouTube video and the dataset's temporal intervals can be accessed from http://www.deepmind.com/kinetics [15].

4.3 Florence 3D

This data was gathered using a stationary Kinect, and only 15 joints were recorded for each body skeleton. It has 215 action sequences with ten subjects and nine actions: wave, sip from a bottle, answer the phone, clap, tighten lace, sit, stand, read, watch, and bow. Some motions, including drinking from a bottle, answering the phone, and reading a watch, are difficult to discern due to a few skeleton joints. To execute leave-one-subject-out cross-validation, conventional experimental settings can be used. The dataset is available at https://www.micc.unifi.it/resources/datasets/florence-3d-actions-dataset [26].

5 Result and interpretation

As stated earlier, very few datasets [15, 26, 28] are available in this domain. Most of the existing state-of-the-art has been carried out with specific datasets, such as the NTU-RGB-D dataset [28]. The study found that none of the existing work has taken all the datasets for either experiment or analysis purposes. In this view, this study is one of its kind that attempts an

Table 3 Performance comparison with Top-1 & Top-5 accuracy on Kinetics skeleton

Existing methods	Top 1	Top 5
Feature Enc. ([12])	14.90%	25.80%
Deep LSTM ([28])	16.40%	35.30%
Temporal Conv. ([17])	20.30%	40.00%
ST-GCN ([51])	30.70%	52.80%
Custom ST-GCN (9 Layer)	31.60%	54.1%
ST-GCN (+eg) ([47])	31.90%	54.1%
Custom ST-GCN (11 Layer)	**32.30%**	**54.50%**

analysis to give a fairly more general comparison of methods using all three datasets [15, 26, 28]. Furthermore, this section evaluates the proposed Custom ST-GCN's performance with two variants i.e. Custom ST-GCN (9-Layer) & Custom ST-GCN (11-Layer) for skeleton-based action recognition with three datasets as NTU-RGB-D, Kinetics-skeleton & Florence 3D. The experiment results were compared with the state-of-the-art methods and summarised in Tables 3, 4 and 5.

The proposed Spatio-temporal graph convolution network method outperforms existing graph-based and LSTM-based algorithms [20, 50]. In addition, unlike other LSTM-based approaches that describe sequence dynamics by revising LSTM, the presented approach incorporates convolution filtering success into recursive learning with a theoretical guarantee [23, 25].

Experimental evaluation has been conducted through Top-1, Top-5 accuracy & mean loss as described in the Eqs. (13), (14), and (15) respectively, which measures the underlying model performance for action classification. Figure 11 depicts the training accuracy vs epoch curve of NTU-RGB-D, Kinectics-skeleton and Florence-3D datasets up to 80 epochs. Figure 12 depicts the performance comparison using the Top-1 accuracy with the NTU-RGB-D dataset. Top-1 accuracy is predicted because the highest probability model likelihood must match the actual value. Figure 13 presents results using the top-5 accuracy with NTU-RGB-D. Top-5 accuracy refers to the model's top-5 highest levels of accuracy compared to the target value. The Figs. 12 and 13 shows a better accuracy than all previous, i.e. 82.7% & 98.1% as Top-1 & Top-5 accuracy, respectively in X-Subject dataset. While, In terms of X-View, it offers better performance than earlier methods, i.e. 90.2% & 99.3% as Top-1 & Top-5 accuracy,

Table 4 Performance comparison with accuracy on Florence 3D

Existing methods	Accuracy
Multi-part Bag-of-Poses ([26])	82.00%
Riemannian Manifold ([8])	87.04%
Lie Group ([34])	90.88%
Graph-Based ([22])	91.63%
MIMTL ([50])	95.29%
P-LSTM ([28])	95.35%
ST-GCN([51])	96.25%
Custom ST-GCN (9 Layer)	97.60%
Custom ST-GCN (11 Layer)	**98.10%**

Table 5 Custom ST-GCN (11-Layer) model performance with datasets

Parameter		Top 1	Top 5	Mean Loss
NTU-RGB-D	X-view	90.20%	99.3%	0.33
	X-Sub	82.70%	98.10%	0.61
Kinetics Skeleton		32.30%	54.50%	3.01
Florence 3D		98.10%	98.70%	0.087

respectively.

$$A-1=\frac{Y^{'}}{Y}*100 \tag{13}$$

Where, $A-1$= Top-1 accuracy, Y'= Number of single correct label prediction, Y= total number of prediction

$$A-5=\frac{X^{'}}{X}*100 \tag{14}$$

Where, $A-5$= Top-5 accuracy, X'= Number of five pair correct label prediction, X= total number of prediction

$$l=\frac{1}{n}\sum_{i=1}^{n}(P-P^{'}) \tag{15}$$

Where, l= mean loss, P'= Predicted label, P= Correct prediction prediction

Table 3 compares the existing model with the proposed model, i.e., Custom ST-GCN on the kinetics skeleton dataset. The proposed model works efficiently in comparison to all other

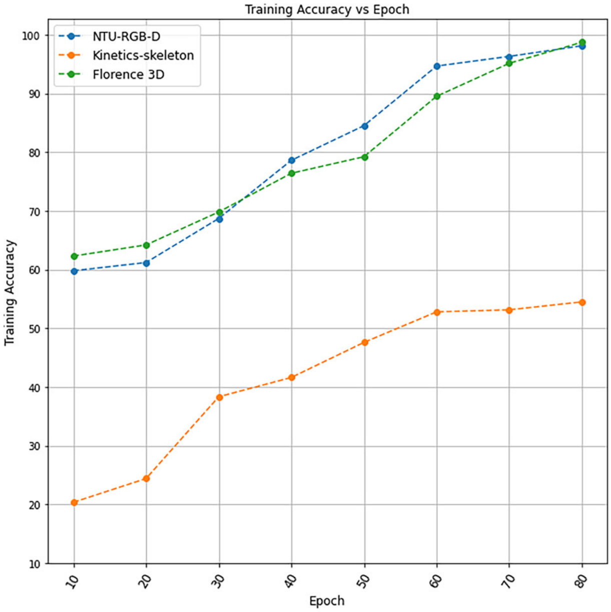


Fig. 11 The training accuracy vs epoch comparison

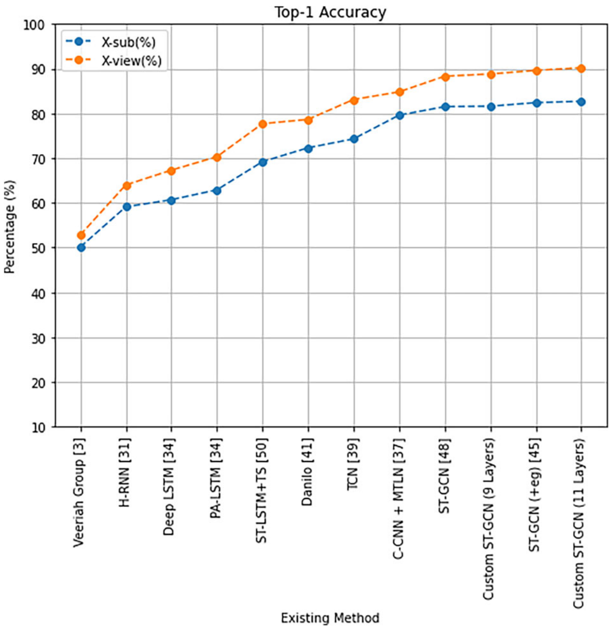


Fig. 12 Comparison in Top-1 accuracy of proposed custom ST-GCN on NTU-RGB-D

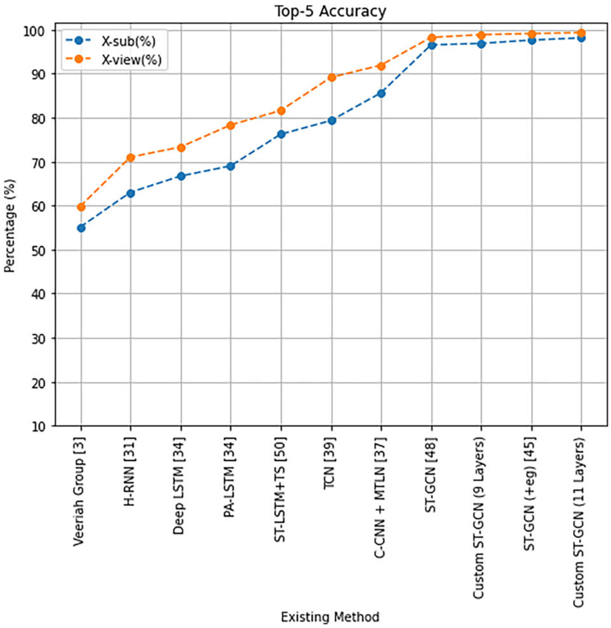


Fig. 13 Performance with Top-5 accuracy of proposed custom ST-GCN on NTU-RGB-D

existing models. It offers a remarkable accuracy than earlier methods, i.e. 32.30% & 54.50% as Top-1 & Top-5 accuracy, respectively.

Table 4 compared the state-of-the-art and proposed custom ST-GCN on the Florence 3D skeleton dataset. The proposed model works efficiently in comparison to all other existing models. It gives significantly higher accuracy than all previous, i.e., 98.10%. In other words, the custom ST-GCN model hikes the performance by 1.92% from one of the top existing performances, i.e., ST-GCN [51].

Table 5 compares the performance characteristics of the presented custom ST-GCN model with NTU-RGB-D, Kinetics, and Florence 3D datasets. Top-1, Top-5 accuracy and mean loss are performance measures used for evaluation. As discussed, the NTU-RGB-D dataset offers two benchmarks: cross-subject (X-Sub), which yields 82.7% as Top-1 accuracy, 98.1% as Top-5 accuracy, and 0.61 as a mean loss. Cross-view (X-View) achieves 90.2% as Top-1 accuracy, 99.3% as Top-5 accuracy and 0.33 as a mean loss. The Top-1 and Top-5 accuracy in the Kinetics skeleton dataset is 32.30%, and 54.50%, respectively & the mean loss is 3.01. The proposed method also gives remarkable results in the Florence 3D dataset, with Top-1 and Top-5 accuracy of 98.10% and 98.70%, respectively. The mean loss is 0.087.

The conventional approach for human activity recognition with RGB-based videos involves splitting the video into frames, treating each frame as a separate image, and feeding these images as input for a deep learning model. The processing of the image itself demands high computation. In contrast, the proposed method extracts the skeleton information from each frame which is numerical values such as joint positions and angles, in a text or CSV file. This approach offers less complexity of the model compared to the traditional method. Also, custom ST-GCN outperforms complexity with a Gflops of 17.5 compared to the other existing method. Similarly, the proposed model's inference time requires less than other cutting-edge techniques. Additionally, skeleton information addresses challenges suffered by the RGB approach, such as low light, occlusion, background clutter, and human pose variation, which can affect the accuracy of human activity recognition tasks.

Table 6 summarises the comparison with respect to GFLOPS & Inference time. FLOPS (floating-point operations per second) is a common metric for measuring the computational complexity and performance of deep learning models. As per the Eq. (16), GFLOPS (billions of FLOPS) is a unit used to express the magnitude of FLOPS. While the Eq. (17) depicts the inference time, which refers to the time to make a prediction by a trained model.

$$GFLOPS = \frac{(No.of FLOPS \times batchsize)}{(inferencetime \times 10^9)} \tag{16}$$

$$InferenceTime = \frac{No.of FLOPS}{(batchsize \times 10^9)} \tag{17}$$

Table 6 Complexity comparison with GFLOPS & Inference Time of ST-GCN vs Custom ST-GCN

Existing method	GFLOPS	Inference Time(s)
ST-GCN ([51])	13.72	0.0251
Custom ST-GCN (11-Layer)	**17.5**	**0.0213**

6 Conclusion and future Work

This research attempts to offer a custom ST-GCN model which outperforms the existing methods. It believes that temporal and spatial features play a vital role in better learning the skeleton and joining the movement. Hence, these features have been used in graph convolutions networks. The combination of spatial-temporal features improves in performance of the ST-GCN model for action classification. This study also presents a fairly general experimental analysis first time by considering all three datasets as NTU-RGB-D, Kinetics-skeleton & Florence 3D. The presented graph-based topologies capture the changing aspects of a motion-based skeleton sequence better than some of the other approaches.

The presented model is limited to modelling long-term temporal dependencies because it only captures local temporal patterns within a short time window. It can limit its use to a bit long time window (long-term dependencies between frames). Furthermore, it is sensitive to the choice of hyperparameters, such as the number of convolutional filters and the filter size. These hyperparameters can significantly impact the model's performance, and finding the optimal values can be challenging.

The ST-GCN model's versatility also offers many possibilities for future research. The challenge of including contextual information like locations, objects, and interactions into ST-GCN. A research direction may be applied in two ways; Firstly, use the proposed approach to multi-user settings, which necessitates the ability to extract several individual skeletons from the image and accurately track them throughout the action sequence. Secondly, the suggested architecture may be combined with algorithms to extract human posture from RGB and depth data. In the future, incorporating an attention mechanism can be a promising approach to improve human activity recognition tasks as it allows the model to focus on the most relevant features and parts of the input sequence. Furthermore, graph structures can capture the relationships between different parts of the input sequence, such as the dependencies between various body parts in a human activity recognition task. These techniques could be exciting directions for future work.

Data availability The data that support the findings of this study are openly available and cited/reference in text.

Declarations

Conflict of interest I declare on behalf of the author that there is not any conflict of interest, either non-financial or commercial among the author.

References

1. Agahian S, Negin F, Köse C (2020) An efficient human action recognition framework with pose-based spatiotemporal features. Engineering Science and Technology, an International Journal 23(1):196–203
2. Al-Janabi S, Salman AH (2021) Sensitive integration of multilevel optimization model in human activity recognition for smartphone and smartwatch applications. Big Data Mining and Analytics 4(2):124–138
3. Avola D, Cinque L, Foresti GL, Marini MR (2019) An interactive and low-cost full body rehabilitation framework based on 3D immersive serious games. J Biomed Inform 89:81–100
4. Cao X, Kudo W, Ito C, Shuzo M, Maeda E (2019) Activity recognition using ST-GCN with 3D motion data. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, pp 689–692

5. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7291–7299
6. Chen X, Koskela M (2015) Skeleton-based action recognition with extreme learning machines. Neurocomputing 149:387–396
7. Chunduru V, Roy M, Chittawadigi RG et al (2021) Hand tracking in 3D space using mediapipe and PNP method for intuitive control of virtual globe. In: 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), pp 1–6. IEEE
8. Devanne M, Wannous H, Pala P, Berretti S, Daoudi M, Del Bimbo A (2015) Combined shape analysis of human poses and motion units for action segmentation and recognition. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 7, pp 1–6. IEEE
9. Dhiman C, Vishwakarma DK (2020) View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Trans Image Process 29:3835–3844
10. Dhiman C, Vishwakarma DK, Agarwal P (2021) Part-wise spatio-temporal attention driven CNN-based 3d human action recognition. ACM Trans Multimed Comput Commun Appl 17(3):1–24
11. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1110–1118
12. Fernando B, Gavves E, Oramas JM, Ghodrati A, Tuytelaars T (2015) Modeling video evolution for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5378–5387
13. Guan Y, Plötz T (2017) Ensembles of deep LSTM learners for activity recognition using wearables. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1(2):1–28
14. Hbali Y, Hbali S, Ballihi L, Sadgal M (2018) Skeleton-based human activity recognition for elderly monitoring systems. IET Comput Vision 12(1):16–26
15. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P et al (2017) The kinetics human action video dataset. arXiv preprint arXiv:1705.06950
16. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3288–3297
17. Kim TS, Reiter A (2017) Interpretable 3D human action analysis with temporal convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 20–28
18. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD (2017) Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 156–165
19. Li C, Cui Z, Zheng W, Xu C, Yang J (2018) Spatio-temporal graph convolution for skeleton based action recognition. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence
20. Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal LSTM with trust gates for 3D human action recognition. In: European Conference on Computer Vision, pp 816–833. Springer
21. Mukherjee S, Anvitha L, Lahari TM (2020) Human activity recognition in RGB-D videos by dynamic images. Multimedia Tools and Applications 79(27):19787–19801
22. Nie F, Wang X, Jordan M, Huang H (2016) The constrained Laplacian rank algorithm for graph-based clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30
23. Patel AS, Vyas R, Vyas OP, Ojha M, Tiwari V (2022) Motion-compensated online object tracking for activity detection and crowd behavior analysis. In: The Visual Computer, pp 1–21
24. Patnaik SK, Babu CN (2021) Bhave M (2021) Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks. Big Data Mining and Analytics 4(4):279–297
25. Pawar K, Jalem RS, Tiwari V (2019) Stock market price prediction using LSTM RNN. In: Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, pp 493–503. Springer
26. Seidenari L, Varano V, Berretti S, Del Bimbo A, Pala P (2013) Weakly aligned multi-part bag-of-poses for action recognition from depth cameras. In: International Conference on Image Analysis and Processing, pp 446–455. Springer
27. Setiawan F, Yahya BN, Chun SJ, Lee SL (2022) Sequential inter-hop graph convolution neural network (SIHGCN) for skeleton-based human action recognition. Expert Syst Appl 195:116566
28. Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1010–1019
29. Singh T, Vishwakarma DK (2021) A deeply coupled convnet for human activity recognition using dynamic and rgb images. Neural Comput Appl 33:469–485

30. Snoun A, Jlidi N, Bouchrika T, Jemai O, Zaied M (2021) Towards a deep human activity recognition approach based on video to image transformation with skeleton data. Multimedia Tools and Applications 80(19):29675–29698
31. Song S, Lan C, Xing J, Zeng W, Liu J (2017) 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
32. Tania S, Rowaida R (2016) A comparative study of various image filtering techniques for removing various noisy pixels in aerial image. International Journal of Signal Processing, Image Processing and Pattern Recognition 9(3):113–124
33. Veeriah V, Zhuang N, Qi GJ (2015) Differential recurrent neural networks for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 4041–4049
34. Vemulapalli R, Arrate F, Chellappa R (2014) 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, pp 588–595
35. Vishwakarma DK, Kapoor R (2012) Simple and intelligent system to recognize the expression of speech-disabled person. In: 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), pp 1–6. IEEE
36. Vishwakarma DK, Dhiman C (2019) A unified model for human activity recognition using spatial distribution of gradients and difference of gaussian kernel. Vis Comput 35(11):1595–1613
37. Vishwakarma DK, Jain K (2022) Three-dimensional human activity recognition by forming a movement polygon using posture skeletal data from depth sensor. ETRI J 44(2):286–299
38. Vishwakarma DK, Kapoor R (2015) Integrated approach for human action recognition using edge spatial distribution, direction pixel and-transform. Adv Robot 29(23):1553–1562
39. Vishwakarma DK, Kapoor R (2017) An efficient interpretation of hand gestures to control smart interactive television. International Journal of Computational Vision and Robotics 7(4):454–471
40. Vishwakarma DK, Singh K (2016) Human activity recognition based on spatial distribution of gradients at sublevels of average energy silhouette images. IEEE Transactions on Cognitive and Developmental Systems 9(4):316–327
41. Vishwakarma DK, Dhiman A, Maheshwari R, Kapoor R (2015) Human motion analysis by fusion of silhouette orientation and shape features. Procedia Computer Science 57:438–447
42. Vishwakarma DK, Rawat P, Kapoor R (2015) Human activity recognition using Gabor wavelet transform and Ridgelet transform. Procedia Computer Science 57:630–636
43. Vishwakarma DK, Kapoor R, Dhiman A (2016) A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics. Robot Auton Syst 77:25–38
44. Vishwakarma DK, Kapoor R, Dhiman A (2016) Unified framework for human activity recognition: an approach using spatial edge distribution and r-transform. AEU-International Journal of Electronics and Communications 70(3):341–353
45. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Frontiers in Robotics and AI 2:28
46. Wan S, Qi L, Xu X, Tong C, Gu Z (2020) Deep learning models for real-time human activity recognition with smartphones. Mobile Networks and Applications 25(2):743–755
47. Wang Q, Zhang K, Asghar MA (2022) Skeleton-based ST-GCN for human action recognition with extended skeleton graph and partitioning strategy. IEEE Access 10:41403–41410
48. Wang X, ZhaoJ, Zhu L, Zhou X, Li Z, Feng J, Deng C, Zhang Y (2021) Adaptive multi-receptive field spatial-temporal graph convolutional network for traffic forecasting. In: 2021 IEEE Global Communications Conference (GLOBECOM), pp 1–7. IEEE
49. Yadav A, Vishwakarma DK (2020) A unified framework of deep networks for genre classification using movie trailer. Appl Soft Comput 96:106624
50. Yang Y, Deng C, Gao S, Liu W, Tao D, Gao X (2016) Discriminative multi-instance multitask learning for 3d action recognition. IEEE Trans Multimedia 19(3):519–529
51. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence
52. Yin J, Han J, Wang C, Zhang B, Zeng X (2019) A skeleton-based action recognition system for medical condition detection. In: 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), pp 1–4. IEEE
53. Yu Y, Li M, Liu L, Li Y, Wang J (2019) Clinical big data and deep learning: Applications, challenges, and future outlooks. Big Data Mining and Analytics 2(4):288–305