Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.034862

ARTICLE

# Action Recognition for Multiview Skeleton 3D Data Using NTURGB + D Dataset

**Rosepreet Kaur Bhogal**[1,*] **and V. Devendran**[2]

[1]School of Electronics and Electrical Engineering, Lovely Professional University, Phagwara, 144411, India

[2]School of Computer Science Engineering, Lovely Professional University, Phagwara, 144411, India

*Corresponding Author: Rosepreet Kaur Bhogal. Email: Rosepreet.kaur@lpu.co.in; Rosepreetkaur12@gmail.com

Received: 29 July 2022 Accepted: 21 December 2022 Published: 09 November 2023

## ABSTRACT

Human activity recognition is a recent area of research for researchers. Activity recognition has many applications in smart homes to observe and track toddlers or oldsters for their safety, monitor indoor and outdoor activities, develop Tele immersion systems, or detect abnormal activity recognition. Three dimensions (3D) skeleton data is robust and somehow view-invariant. Due to this, it is one of the popular choices for human action recognition. This paper proposed using a transversal tree from 3D skeleton data to represent videos in a sequence. Further proposed two neural networks: convolutional neural network recurrent neural network_1 (CNN_RNN_1), used to find the optimal features and convolutional neural network recurrent neural network network_2 (CNN_RNN_2), used to classify actions. The deep neural network-based model proposed CNN_RNN_1 and CNN_RNN_2 that uses a convolutional neural network (CNN), Long short-term memory (LSTM) and Bidirectional Long short-term memory (BiLSTM) layered. The system efficiently achieves the desired accuracy over state-of-the-art models, i.e., 88.89%. The performance of the proposed model compared with the existing state-of-the-art models. The NTURGB + D dataset uses for analyzing experimental results. It is one of the large benchmark datasets for human activity recognition. Moreover, the comparison results show that the proposed model outperformed the state-of-the-art models.

## KEYWORDS

Activity; recognition; multiview; LSTM; BiLSTM; NTURGB + D

## 1 Introduction

Action is when we do something, especially when dealing with anything like an object or human. The goal of any human activity recognition system is to recognize ongoing activities from ongoing videos automatically. Recognition of human activities enables real-time monitoring of public places like airports and stations can monitor patients, children and elderly persons [1]. Vision-based activity recognition systems highly impact various motivating application domains, like behavioral biometrics, Content-based video analysis, security and surveillance, interactive application and environment,

animation and synthesis [2]. In behavioral biometrics, various approaches are based on Fingerprint, Face, or Iris and are used to recognize human-based physical or behavioral cues. In this approach, the subject's cooperation is not required and only needed to know the subject's activity. Gait recognition [3] could be the most challenging application area. After all, a person walking characteristics can identify the person through closed-circuit television (CCTV) footage because everyone has a distinct walking style like other biometrics. Today, with fast-growing technology, people can share and search multimedia content, such as images, music and Video. Searching for desired content is very challenging for a retrieval system to find a subset of objects with similar content [4]. Summarizing and retrieving consumer content, such as general activities like sports or cooking videos, are one of the most commercial applications under content-based video analysis. They were developing a visual monitoring system that observes moving objects in a site and learns the pattern of activity from those activities. That system comprises motion tracking, activity classification and event detection.

An area can be significant to observe from a single camera, so many such sensor units use around the site. Cameras are attached to poles, trees and buildings for an outdoor setting. The indoor setting involves attaching to walls and furniture [5]. Intelligent surveillance got more research attention because of effective monitoring of public places, airports, railway stations, shopping malls, crowded places and military installations, or uses intelligent healthcare facilities like fall detection in older people's homes [6]. Often, the motive is to detect, recognize or learn exciting events, defined as suspicious events, irregular behavior, uncommon behavior, unusual activity/event/behavior and abnormal behavior, or anomaly [7].

For such activity, using CCTV cameras to record or observe scenes the user has become ubiquitous. Although recording videos through cameras is cheap, affordable and popular in today's scenario. However, the agents for observing outliers and analyzing the footage are also limited and unreasonable. Wherever video cameras use in the room, they experience poor monitoring due to genuine reasons like the fatigue of the observer. Due to long monitoring hours, the operator can skip noticing suspicious activity, which is generally of short duration. This application comes under security and surveillance because detecting unusual activity at the right time is essential. In interactive applications and environments, the interaction between humans and computers is one of the challenges in designing a human-computer interface. In today's scenario, smart devices are capturing data and analyzing users. The relevant information can extract from the activity tracker for activity recognition. The framework explores fog computing to the cloud for reducing computation proposed in [8]. An interactive domain such as smart rooms responding to a person's gesture can directly or indirectly benefit the user [9]. Such as music according to the user's mood when entering the room. Animation and synthesis require an extensive collection of motions the animator uses to make high-quality animation or movies. Any application can relate human motion to any environment, including training military soldiers, firefighters and other personnel [2,10].

Human activity recognition consists of preprocessing, segmentation, Feature extraction, dimension reduction, and classification. However, various data modalities are available to detect action from activities. Instead of other modalities, 3D information uses to track movement. 3D information contains coordinates value that helps to track body joints efficiently [11]. Nowadays, recording videos at various angles is called multiview data. Multiview learning [12,13] is essential for action recognition. The camera can employ at any angle for recording actions. There is a requirement for the system to detect activities, which can handle many views for identifying actions. Nowadays, deep learning-based methods have accomplished importance in human activity recognition. The recurrent neural networks-based system is considered adequate for sequential data handling [14] and specially designed with LSTM [15], or BiLSTM [16] layered recurrent networks.

It is interesting to know the recent development in activity recognition. The skeleton data has joints describing the body's movement and pose. In multiview action recognition, 3D information can prefer and get more attention. In [17], the hand-crafted feature has been calculated and given to the CNN-based model for skeleton-based action recognition. The work has high computational complexity based on the state-of-the-art comparison. In other words, a convolutional network uses additional features like joint distribution trajectories [18]. Instead of CNNs, recurrent neural networks (RNNs) can prefer because they store the information based on time dependencies which is the essential information in action recognition. In [19] shows multi-model strategies using LSTM. LSTM has proven to be a good choice for data where time-based information is a concern. A few selective frames can choose to find features in the proposed method. Finding features from a smaller number of frames can reduce the computational complexity of the system because frames in videos have replicated activities also. The method [20] used a deep neural network and suggested selecting a keyframe. The LSTM is often applied [21] differential LSTM, which proposes feature enhancement. A densely BiLSTM network is presented in [22], which outperforms spatial and temporal data. In contrast to all these methods, the proposed method in this paper with two neural networks designed with LSTM and BiLSTM layers.

This section includes a survey of the current knowledge, including essential findings based on applications. Providing a machine that can detect or recognize activities from videos through which a human can understand or act based on activity. Main contribution towards action recognition with the proposed work:

1) Technique for preprocessing of 3D skeleton information.
2) The deep neural network can use as a pretrained network for multiview data.
3) Network which can use for feature reduction.
4) Deep neural network for classification of action from 3D skeleton information.
5) This proposed method is independent of multiview and multiple subjects.

The rest paper organizes as follows: Section 2 presents the proposed method. Section 3 includes experimental details. In the last, concluded remarks in Section 4.

## 2 Proposed Method

The NTURGB + D dataset has various data modalities from these modalities, 3D joint information used in the proposed methodology. The information in a skeleton file from which three coordinates, X, Y and Z, have been extracted and used for further representation in the form of optimal features. The steps for extracted 3D data has given in Fig. 1.

The skeleton file has various information, i.e., body ID, Clipped edges, hand left confidence, hand left state, hand right confidence, hand right state, lean X, lean Y, joint count and Joints. Take further on Joints: X, Y, Z, depth X, depth Y, color X, color Y, orientations and tracking details are available. Each frame of the Video has information on X, Y and Z, shown in Fig. 2.

The X, Y and Z are the 3D coordinates values of joints. There are 25 Joints body positions for activity recognition, as given in Fig. 3. The dimension of each vector is $[25 \times 1]$. Each vector is rearranged in the tree so that network can train in a more appropriate form.

The proposed method consists of three significant components illustrated below.

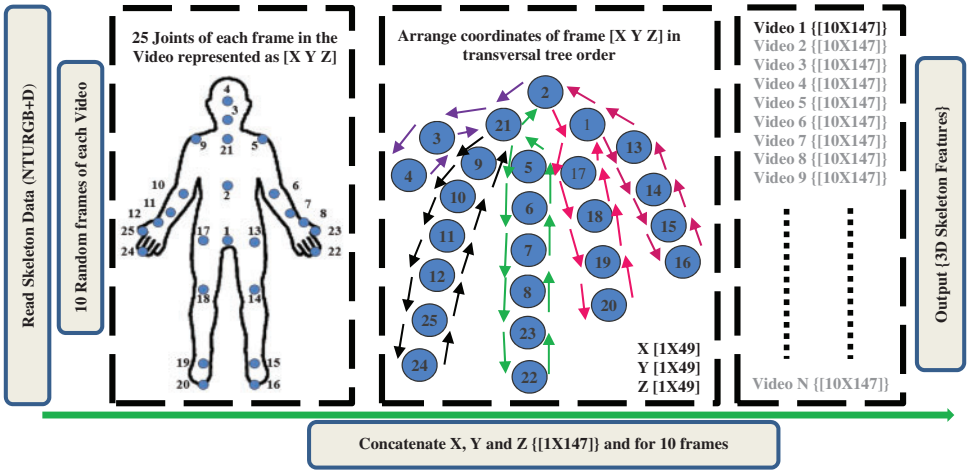


**Figure 1:** 3D skeleton features using the transversal tree

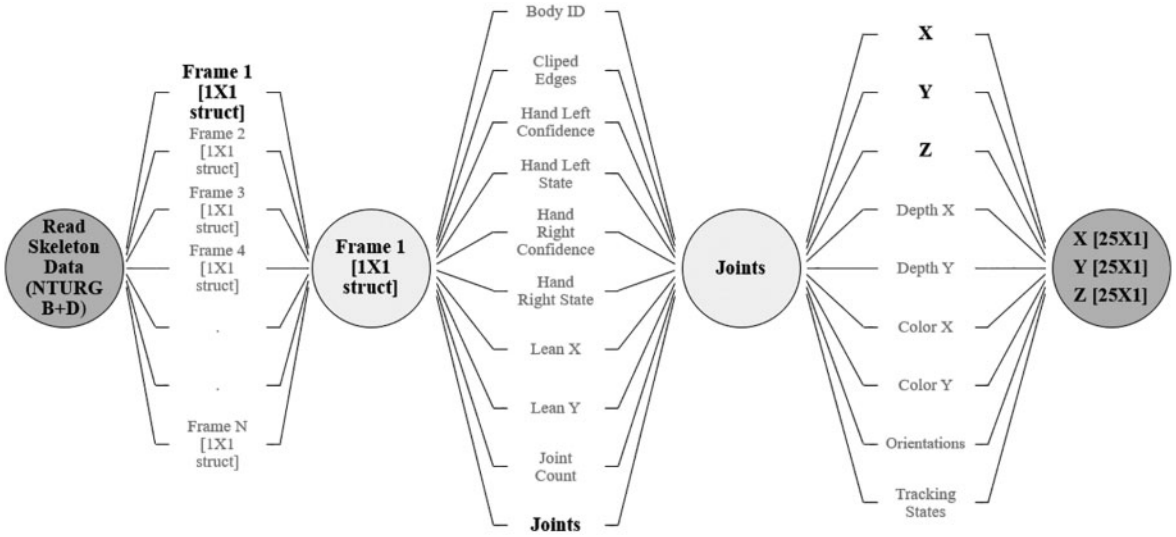


**Figure 2:** Step for 3D coordinates of skeleton 3D NTURGB + D dataset

### *2.1 3D Skeleton Pre-Processing (Representation of 3D Skeleton Data into Sequences Using Transversal Tree)*

For human activity recognition, the features play an essential role. The skeleton sequences of NTURGB + D contain Skeleton 3D data for all videos of 60 classes. Each Video from the dataset has a different number of frames. The random ten frames have been considered for the same size output to make the Skeleton feature. For example, if any video contains 100 frames. Then, the index of frame selection is [1 11 21 31 41 51 61 71 81 91]. Each frame has joint information containing the details of X, Y and Z coordinates values. The X, Y and Z are the 3D coordinates of skeleton sequences. The 25 joints correspond to which point of the human body, as given in Fig. 3. The size of each X, Y and Z is [25 1]. All coordinates are arranged as transversal tree orders [17], in which the index of the joint number has stores accordingly. The order of making a new sequence is [2 21 3 4 3 21 9 10 11 12 25 24 25 12 11 10 9 21 5 6 7 8 23 22 23 8 7 6 5 21 2 1 17 18 19 20 19 18 17 1 13 14 15 16 15 14 13 1 2]. The

total number of joints available is 25. To give the network a sequence with that it can train efficiently. Make that number of the joint index 49 for each X, Y and Z as per the transversal tree. Concatenate the ten frames X, Y and Z, 3D skeleton data and size becomes [10 × 49] where 10 is the number of frames chosen from each Video and 49 is the X, Y and Z data index. There is a requirement to give a pattern for system learning. Joints arrange to give relative motion information of coordinates. The joints arrange in the manner of body parts, i.e., the torso, right arm, left arm, right leg and left leg. Each Video from the dataset has represented a new skeleton 3D of fixed size [10 × 147]. These are the feature which is the input to CNN_RNN_1.

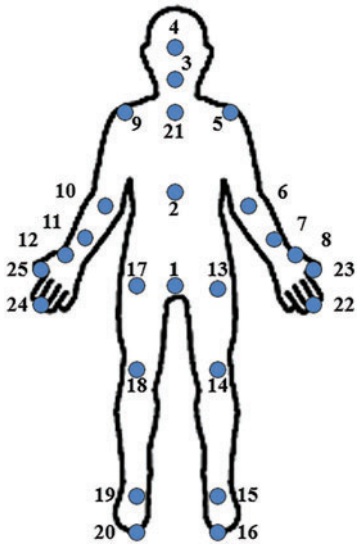


**Figure 3:** Joint location of the human body of skeleton 3D data [25 joints]

### *2.2 CNN_RNN_1 for Optimal Features*

In this work, multiple LSTMs and CNNs uses for human activity recognition. Generally, researchers use a pre-trained network to find optimal features. There is no such pre-trained network that acts optimally for the NTURGB + D dataset. The LSTMs-based model is contextually dependent on the temporal domain. Moreover, CNN-based models focus on spatial information. Temporal and spatial information are essential features for action classification using Videos. CNN_RNN_1 trains to find optimal features for another network called CNN_RNN_2. First, the CNN_RNN_1 trains using a dataset then the network is used to find optimal features. Input videos dataset fix for CNN_RNN_1 as well as CNN_RNN_2 for training purposes. The video dataset for validation and testing is distinct from training for CNN_RNN_1.

The CNN_RNN_1 consists of 20 layers, including LSTMs and CNNs with other essential layers. After giving the input, the first layer is the folding layer, which converts a batch of an image sequence to a collection of images and it converts the sequence for the next layer to do convolution. The three consecutive convolutional layers have 32, 48 and 64 kernels, respectively. After each Convolutional layer, there is a ReLU layer and a Maxpooling layer. The ReLU layer uses to operate any value zero if the value is less than zero.

Furthermore, the Max pooling layer uses to downsample the input and half the number of samples is the output. The unfolding layer restores the sequence data of input data after sequence

folding. After the unfolding layer, a flattened layer converts data into a single column. The flattened sequence passes to LSTM layers. Each LSTM layer ended with a dropout layer, half the sample length. The number of neurons is the same in the LSTM layer, which is 128. The dropout layer with a probability of 0.5 to avoid network overfitting during learning. The last three layers fully connect with the number of neurons, same as the number of classes, i.e., 60. Softmax layer for determining the probability corresponding to each class and classification layer for assigning class as per probability based determined with softmax layer.

### *2.3 CNN_RNN_2 for Action Classification*

CNN_RNN_2 is proposed for the classification and is first required to load the trained proposed CNN_RNN_1. With CNN_RNN_1, optimal features calculate for the same training, validation and testing data used for CNN_RNN_2. The dimension for Skeleton 3D features of each Video is [128 1], as mentioned in Fig. 4. The sequence input gives to the flattened layer. In CNN_RNN_2, the BiLSTM layer uses as in Fig. 5. The two LSTM networks are connected in opposite directions to make BiLSTM. In other words, using the BiLSTM network in our model results in long-term bidirectional relationships, subtracted by going back and forth several times in the vector sequence embedded in all parts of the Video [23]. BiLSTM layer uses to store time dependencies of data and preference for classification of CNN_RNN_2 used here to classify the action from videos.

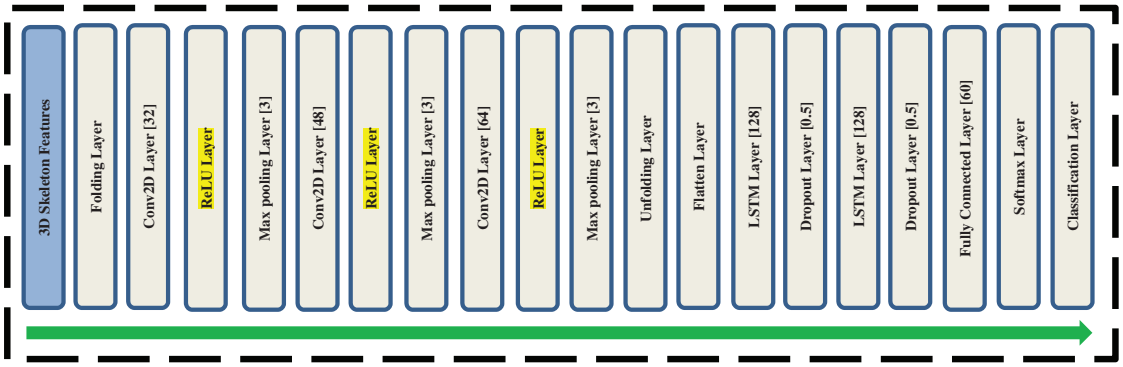


**Figure 4:** Architecture of CNN_RNN_1 for feature reduction

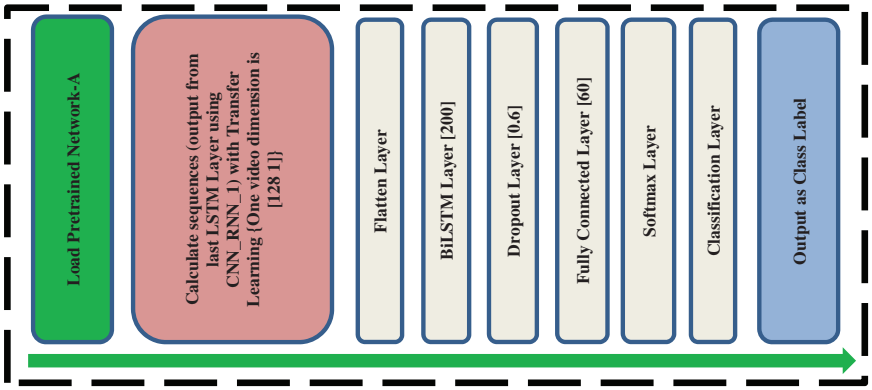


**Figure 5:** Architecture of CNN_RNN_2 for classification

## 3 Experiment

### *3.1 About Dataset (NTURGB + D)*

The proposed method evaluates on dataset NTURGB + D. There are 4 data modalities available, i.e., depth maps, 3D joints information, RGB frames and IR sequences. In this paper, the 3D joints information only use. The joint information consists of 3-dimensional locations of 25 major body joints, as shown in Fig. 2. The major body joints help to detect and track the movement of each body part of the human body. The dataset contains 60 actions classes-based information and 56880 videos recorded using the Microsoft Kinect v2 sensor. There are 40 distinct subjects between the age of 10 to 35 years. The dataset videos recording at three different angles, i.e., 45°, −45° and 0°, as shown in Fig. 6. The activities are divided into three major groups: 40 daily actions, 9 human health-related and 11 mutual actions in the dataset. This dataset has a variation in the number of subjects and ages of subjects [24].

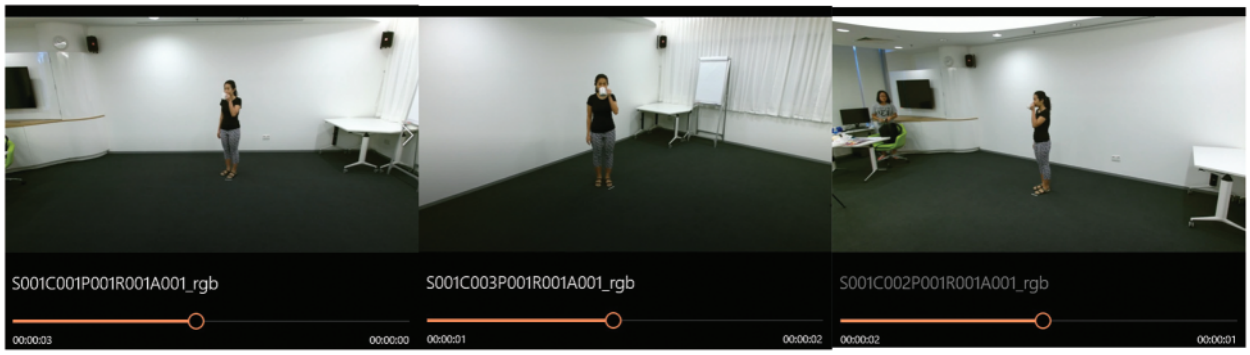


**Figure 6:** Three views in NTURGB + D (side view (+45°), front view (0°) and side view (−45°))

### *3.2 Implementation Details*

The experiment performs on the system with a 1.6 GHz Intel Core i5-4200U, 8 GB RAM and 1 TB SSD running a Windows 10 with the 64-bit operating system. There are CNN_RNN_1 and CNN_RNN_2; both train, validate and test with the dataset. The CNN_RNN_1 has 20 layers which train by using a dataset. No pre-trained network is available for this application to find optimal classification features. The filter size is [3 3] for the convolutional 2D filter. The Adam optimizer uses during training with a learning rate of $10^{-4}$. The minibatch size is 128, with the number of iterations per epoch being 32 upon max epoch 500. The same hyperparameter has been considered for CNN_RNN_2 also. The experiment performs on a total of 960 videos of NTURGB + D. It includes three angled-view videos for all 60 classes.

### *3.3 Results*

The arrangement of joints aims to represent the 3-dimensional coordinates in some pattern. So, networks can learn the pattern in a way; it can identify the activities efficiently. The captured 3D information preprocess for the network. After preprocessing, the 20-layer CNN_RNN_1 network train with some videos. The same CNN_RNN_1 uses to calculate optimal features of which dimension is much less than the input data. It can use as a feature reduction technique to represent a video in a single column of length [128 1]. When all the videos are the same size as [128 1], these features are the

input to the second network called CNN_RNN_2. CNN_RNN_2 is used to classify the action from 3D information. The conclusion is that there are two networks, one for feature reduction technique and another for classification. The experimental results show using two graphs in Figs. 7 and 8. Monitoring the training and validation accuracy progress plot is helpful when training any network. It shows how quickly accuracy is improving and whether a network is starting to overfit with data or not. The training and validation accuracy plot shows that the network is not overfitting or underfitting. Fig. 7 shows the training and validation accuracy plot *vs.* the number of iterations. This designed network is not having any overfitting issue with data. Initially, at zero iteration, accuracy is low. However, after the 100th iteration, the accuracy increases and becomes constant for a few iterations and at max epoch 500th, it ends up with an accuracy of 88.89%.

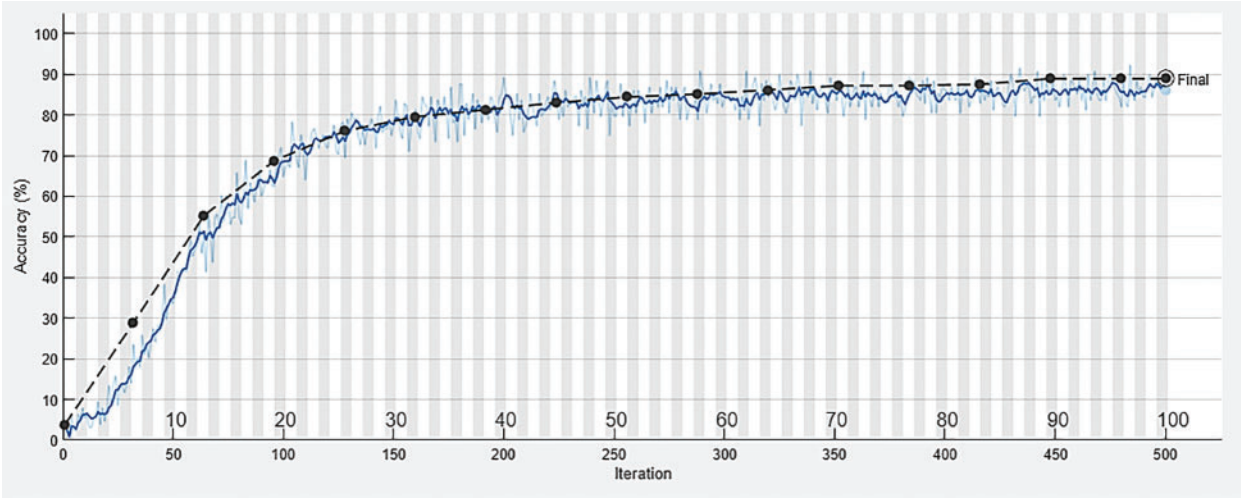


**Figure 7:** Training progress plot (Accuracy *vs*. number of iterations) [Blue line represents training progress and the black line represents validation progress]

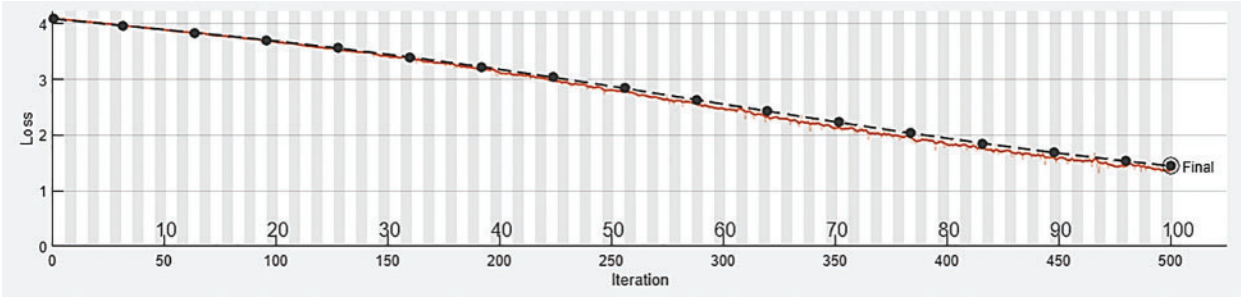


**Figure 8:** Validation loss plot (Loss *vs*. number of iterations) [Red line represents training loss progress and the black line represents validation loss progress]

Fig. 8 shows the training and validation loss function *vs*. the iteration plot. The training and validation losses perfectly fit with the data. However, the loss is more when the batch size is small. As batch size increases, the loss decreases, as given in Fig. 8. The wiggle is minimal when the batch size is the entire dataset because gradient update improves the loss function monotonically. At the maximum epoch, the final validation loss ends up at 1.45. Table 1 compares proposed methodologies with other states of art methods in terms of accuracy. That comparison shows that the proposed method can prefer for action recognition.

**Table 1:** Comparison with other methods based on accuracies

| Method | Accuracy (%) |
|---|---|
| Multidimensional indexing [25] | 84.6 |
| HMM [26] | 88.3 |
| PCA and HMM [27] | 87.5 |
| Memory-based attention control system [28] | 80 |
| Dynamic time warping [29] | 80.05 |
| Bipartite graph [30] | 82.8 |
| Multi-sensor fusion [31] | 88 |
| Deep learning-based hierarchical feature model [32] | 70.32 |
| Deep convolutional neural network [33] | 41.5 |
| Collaborative sparse coding [34] | 79.18 |
| Convex multiview semi-supervised classification [35] | 59.08 |
| Scene flow to action map and ConvNets [36] | 61.94 |
| Convolutional neural network [37] | 66.29 |
| Multiview fusion [14] | 85.9 |
| Graph convolutional networks [38] | 88.2 |
| **Proposed method** | **88.89** |

Fig. 9 shows the confusion matrix for 60 classes where each row instance depicts actual classes and each column as predicted classes. The diagonal green colored boxes show the correct number of classes identified. For example, considering the first row and column, the class label with A001 is the same in 6 videos. However, one Video of A001 recognizes as A003. Likewise, the whole matrix can understand. Table 2 shows the class label with the activity name as recognized actions.


**Figure 9:** Confusion matrix (Actual class *vs.* predicted class)

**Table 2:** Activity ID with activity name

| Activity_ID | Name of activity | Activity_ID | Name of activity | Activity_ID | Name of activity | Activity_ID | Name of activity |
|---|---|---|---|---|---|---|---|
| A001 | Drink water | A016 | Put on a shoe | A031 | Point to something | A046 | Back pain |
| A002 | Eat meal | A017 | Take off a shoe | A032 | Taking a selfie | A047 | Neck pain |
| A003 | Brush teeth | A018 | Put on glasses | A033 | Check time (from watch) | A048 | Nausea/ vomiting |
| A004 | Brush hair | A019 | Take off glasses | A034 | Rub two hands | A049 | Fan self |
| A005 | Drop | A020 | Put on a hat/cap | A035 | Nod head/bow | A050 | Punch/slap |
| A006 | Pick up | A021 | Take off a hat/cap | A036 | Shake head | A051 | Kicking |
| A007 | Throw | A022 | Cheer up | A037 | Wipe face | A052 | Pushing |
| A008 | Sit down | A023 | Hand waving | A038 | Salute | A053 | Pat on back |
| A009 | Stand up | A024 | Kicking something | A039 | Put palms together | A054 | Point finger |
| A010 | Clapping | A025 | Reach into pocket | A040 | Cross hands in front | A055 | Hugging |
| A011 | Reading | A026 | Hopping | A041 | Sneeze/cough | A056 | Giving object |
| A012 | Writing | A027 | Jump up | A042 | Staggering | A057 | Touch pocket |
| A013 | Tear up paper | A028 | Phone call | A043 | Falling down | A058 | Shaking hands |
| A014 | Put on jacket | A029 | Play with phone/tablet | A044 | Headache | A059 | Walking towards |
| A015 | Take off jacket | A030 | Type on a keyboard | A045 | Chest pain | A060 | Walking apart |

## 4 Conclusion

This paper presents multiview-based skeleton action recognition using deep neural networks. This paper proposes the networks, i.e., CNN_RNN_1 and CNN_RNN_2, where CNN_RNN_1 uses for feature reduction technique and CNN_RNN_2 for action classification. The activities classification uses 3D skeleton information of all three views from the dataset NTURGB + D for 60 classes. The designed system outperforms all the other state-of-the-art methods. The accuracy of action can improve by including more layers in the network. The system can also design with two-stream or three-stream input networks to improve evaluation parameters. This work will extend by developing two-stream networks.

**Acknowledgement:** The authors are thankful to all members of the Rapid-Rich Object Search (ROSE) Lab from Nanyang Technological University, Singapore and Peking University, China. Their vision is to create the most extensive collection of datasets. They have provided an opportunity to use an action recognition dataset.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** The authors confirm contribution to the paper as follows: Literature review and problem identification: Rosepreet Kaur Bhogal, V. Devendran; Data collection: Rosepreet Kaur Bhogal; Analysis and interpretation of results: Rosepreet Kaur Bhogal, V. Devendran; Draft manuscript preparation: Rosepreet Kaur Bhogal, V. Devendran. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The members of the Rapid-Rich Object Search (ROSE) Lab from Nanyang Technological University, Singapore and Peking University, China has created the most extensive collection of datasets. The researcher can request the dataset on website https://rose1.ntu.edu.sg/dataset/actionRecognition/.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] B. V. Rani and P. Singh, "A survey on electronic health records (EHRS): Challenges and solutions," in *2022 6th Int. Conf. on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 655–658, 2022.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[3] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother *et al.,* "The human ID gait challenge problem: Data sets, performance and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[4] M. Ramezani and F. Yaghmaee, "A review on human action analysis in videos for retrieval applications," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 485–514, 2016.

[5] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[6] A. Ahmed, M. M. Khan, P. Singh, R. S. Batth and M. Masud, "IoT-based real-time patients vital physiological parameters monitoring system using smart wearable sensors," *Neural Computing and Applications*, vol. 34, no. 22, pp. 19397–19673, 2022.

[7] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—A review," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, 2012.

[8] A. Kaur, P. Singh and A. Nayyar, "Fog computing: Building a road to IoT with fog analytics," *Studies in Big Data*, vol. 76, pp. 59–78, 2020.

[9] P. Singh, A. Kaur and N. Kumar, "A reliable and cost-efficient code dissemination scheme for smart sensing devices with mobile vehicles in smart cities," *Sustainable Cities and Society*, vol. 62, pp. 102374, 2020.

[10] F. R. Khan, M. Muhabullah, R. Islam, M. M. Khan, M. Masud *et al.,* "A cost-efficient autonomous air defense system for national security," *Security and Communication Networks*, vol. 2021, 9984453, 2021.

[11] C. Li, Y. Hou, P. Wang and W. Li, "Multiview-based 3D action recognition using deep networks," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 1, pp. 95–104, 2019.

[12] S. Sun, "A survey of multiview machine learning," *Neural Computing and Applications*, vol. 23, no. 7–8, pp. 2031–2038, 2013.
[13] C. Hong, J. Yu, J. Wan, D. Tao and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5659–5670, 2015.
[14] Z. Fan, X. Zhao, T. Lin and H. Su, "Attention-based multiview re-observation fusion network for skeletal action recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 363–374, 2019.
[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[16] R. L. Abduljabbar, H. Dia and P. W. Tsai, "Unidirectional and bidirectional LSTM models for short-term traffic prediction," *Journal of Advanced Transportation*, vol. 2021, pp. 5589075.1–5589075.16, 2021.
[17] Y. Du, Y. Fu and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conf. on Pattern Recognition (ACPR)*, Nanjing, China, pp. 579–583, 2016.
[18] Y. Hou, Z. Li, P. Wang and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, 2018.
[19] J. Liu, A. Shahroudy, D. Xu and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," *Lecture Notes in Computer Science*, vol. 9907, 2016.
[20] H. H. Phan, T. T. Nguyen, N. H. Phuc, N. H. Nhan, D. M. Hieu *et al.,* "Key frame and skeleton extraction for deep learning-based human action recognition," in *2021 RIVF Int. Conf. on Computing and Communication Technologies*, Hanoi, Vietnam, pp. 1–6, 2021.
[21] K. Hu, F. Zheng, L. Weng, Y. Ding and J. Jin, "Action recognition algorithm of spatio–temporal differential LSTM based on feature enhancement," *Applied Sciences*, vol. 11, no. 17, pp. 7876, 2021.
[22] J. Y. He, X. Wu, Z. Q. Cheng, Z. Yuan and Y. G. Jiang, "DB-LSTM: Densely-connected bi-directional LSTM for human action recognition," *Neurocomputing*, vol. 444, pp. 319–331, 2021.
[23] X. Wu and Q. Ji, "TBRNet: Two-stream BiLSTM residual network for video action recognition," *Algorithms*, vol. 13, no. 7, pp. 1–21, 2020.
[24] A. Shahroudy, T. T. Ng, Q. Yang and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2123–2129, 2016.
[25] J. B. Arie, Z. Wang, P. Pandit and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1091–1104, 2002.
[26] F. Niu and M. A. Mottaleb, "View-invariant human activity recognition based on shape and motion features," in *IEEE Sixth Int. Symp. on Multimedia Software Engineering*, Miami, USA, pp. 546–556, 2004.
[27] M. Ahmad and S. W. Lee, "HMM-based human action recognition using multiview image sequences," in *18th Int. Conf. on Pattern Recognition (ICPR'06)*, Milan, Italy, pp. 263–266, 2006.
[28] K. F. MacDorman, H. Nobuta, S. Koizumi and H. Ishiguro, "Memory-based attention control for activity recognition at a subway station," *IEEE Multimedia*, vol. 14, no. 2, pp. 38–49, 2007.
[29] S. Cherla, K. Kulkarni, A. Kale and V. Ramasubramanian, "Towards fast, view-invariant human action recognition," in *2008 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, Nashville, USA, pp. 1–8, 2008.
[30] J. Liu, M. Shah, B. Kuipers and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Conf. on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 3209–3216, 2011.
[31] F. Zhu, L. Shao and M. Lin, "Multiview action recognition using local similarity random forests and sensor fusion," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 20–24, 2013.
[32] M. Hasan and A. K. R. Chowdhury, "A continuous learning framework for activity recognition using deep hybrid feature models," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1909–1922, 2015.
[33] H. Zhu, J. B. Weibel and S. Lu, "Discriminative multi-modal feature fusion for RGBD indoor scene recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 2969–2976, 2016.

[34] W. Wang, Y. Yan, L. Zhang, R. Hong and N. Sebe, "Collaborative sparse coding for multiview action recognition," *IEEE Multimedia*, vol. 23, no. 4, pp. 80–87, 2016.
[35] F. Nie, J. Li and X. Li, "Convex multiview semi-supervised classification," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5718–5729, 2017.
[36] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang *et al.,* "Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 416–425, 2017.
[37] A. A. Liu, N. Xu, W. Z. Nie, Y. T. Su and Y. D. Zhang, "Multi-domain and multi-task learning for human action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 853–867, 2019.
[38] J. Xie, W. Xin, R. Liu, L. Sheng, X. Liu *et al.,* "Cross-channel graph convolutional networks for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 9055–9065, 2021.