I.J. Information Technology and Computer Science, **2025, 6, 176-193**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.06.10

Lightweight 3DCNN-BiLSTM Model for Human Activity Recognition using Fusion of RGBD Video Sequences

Vijay Singh Rana*
Department of Computer Applications, College of Smart Computing Roorkee, COER University Roorkee India
E-mail: vijayranasrb25@gmail.com
ORCID iD: https://orcid.org/ 0009-0006-0234-2225
*Corresponding author

Ankush Joshi
Department of Computer Science and Applications, College of Smart Computing, COER University Roorkee India
E-mail: ankushjoshi1987@gmail.com
ORCID iD: https://orcid.org/ 0000-0003-0873-3340

Kamal Kant Verma
School of Computer Science and Engineering, IILM University Greater Noida, India
E-mail: dr.kamalverma83@gmail.com
ORCID iD: https://orcid.org/ 0000-0003-0030-3688

Received: 11 June 2025; Revised: 26 August 2025; Accepted: 23 October 2025; Published: 08 December 2025

Abstract: Over the past two decades, the automatic recognition of human activities has been a prominent research field. This task becomes more challenging when dealing with multiple modalities, different activities, and various scenarios. Therefore, this paper addresses activity recognition task by fusion of two modalities such as RGB and depth maps. To achieve this, two distinct lightweight 3D Convolutional Neural Network (3DCNN) are employed to extract space time features from both RGB and depth sequences separately. Subsequently, a Bidirectional LSTM (Bi-LSTM) network is trained using the extracted spatial temporal features, generating activity score corresponding to each sequence in both RGB and depth maps. Then, a decision level fusion is applied to combine the score obtained in the previous step. The novelty of our proposed work is to introduce a lightweight 3DCNN feature extractor, designed to capture both spatial and temporal features form the RGBD video sequences. This improves overall efficiency while simultaneously reducing the computational complexity. Finally, the activities are recognized based the fusion scores. To assess the overall efficiency of our proposed lightweight-3DCNN and BiLSTM method, it is validated on the 3D benchmark dataset UTKinectAction3D, achieving an accuracy of 96.72%. The experimental findings confirm the effectiveness of the proposed representation over existing methods.

Index Terms: Human Activity Recognition, Lightweight 3DCNN, Bidirectional LSTM, Decision Level Fusion, Depth Map.

1. Introduction

Human activity Recognition (HAR) in video sequences has an active research area from past two decades in computer vision due to its wide variety of applications and acceptance in diverse domains, including healthcare, daily living activities, intelligent monitoring, security and assistance in disability or monitoring elderly person. There is a lot of ongoing research happening in the field. The research work presented in this paper focuses on daily living activity recognition system specially designed to monitor human actions automatically in video sequences. Due the readily availability of video cameras in our day to day life, HAR has numerous applications in various domain such as video analysis, human to computer interactions, robotic, healthcare, security and crime investigations etc. Essentially, in video data HAR interpret two aspects spatial and temporal. The spatial information identifies the objects, context and visual

scenarios available in the specific frame, however the temporal data gives the time information of the action between the consecutive frames. The temporal dynamics between the consecutive frames is more valuable for action recognition. Here, the process becomes more computationally complicated because a video sequence might consist of hundreds of frames that needs to be process independently. Despite its potential benefits, still HAR faces substantial challenges in action modeling, accuracy and robustness when dealing only with one type of information such as RGB videos to map the complexities of the real work scenarios [1]. The spatial and temporal information are more appropriate and commonly used for action recognition task [2]. Recently, the widespread availability of low budgeted RGBD depth cameras such as Microsoft Kinect, Xbox 360 and ASUS Xtion Kinect motivated the research community towards efficient and low budget solutions, significantly simplifying the challenges of HAR at a certain level. The research has now inclined towards learning the activities using these low-cost depth cameras [3,4].

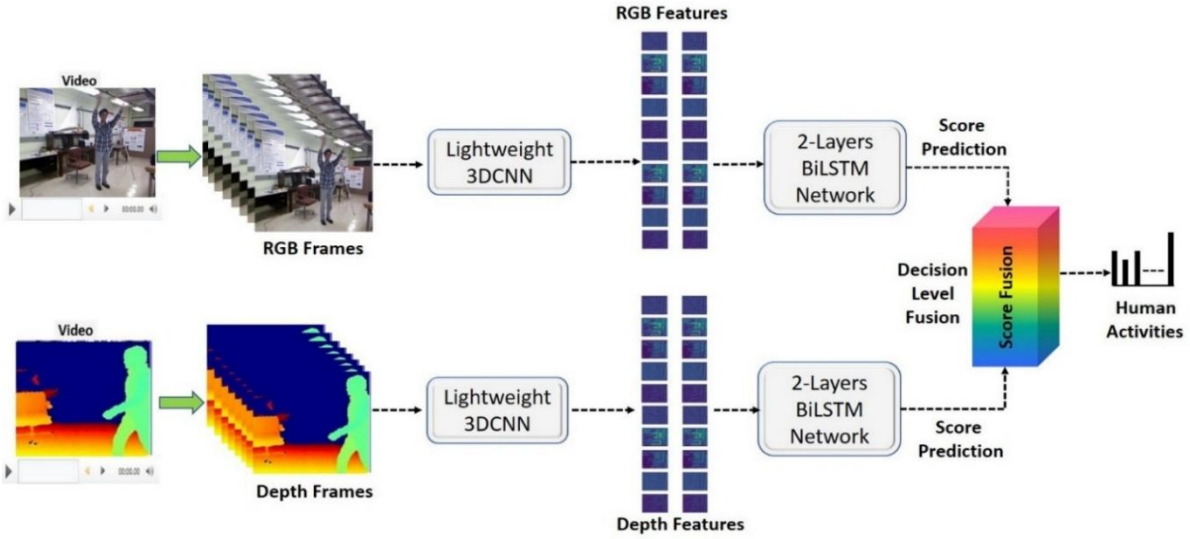

Fig.1. Flow diagram of proposed system

The depth camera has a number of advantages over the RGB, including the fact that it is not affected by changes in illumination, variable lightening condition, or moving background etc. The row data capture by depth camera are insensitive to the cluttered background and occlusion. Additionally, it provides an extra multi dimensionality data modality such as 3d skeleton for improved action recognition. Zhu et al. [5]

Furthermore, the challenging task is to learn the temporal variation from the RGBD video sequences and could become more difficult especially when activity recognition is being performed by real time surveillance system or through online mode. For instance, conventional trajectory-based method relies heavily on hand crafted and optical flow features [6,7]. Likewise, some deep end-to-end multi stream method [8,9] utilize multiple 2D networks that corresponds to appearance and optical flow. While these methods deliver good performance, obtaining optical flow features is computationally intensive, making them less practical for large-scale datasets and real-time surveillance systems. To address with the issue of expensive spatial and temporal computations a lightweight 3DCNN has been employed in this research work. A 3DCNN has the capability to directly learns the spatial and temporal variations, greatly enhancing the performance of classification and reducing computational time. A variant of 3DCNN such as lightweight CNN is becoming more popular for special temporal feature extraction from the input RGBD video sequences. For instant, Ullah et al. [10] proposed a lightweight deep learning assisted framework for activity recognition. First, they detected human and tracked in the entire video stream using minimum output sum of square (MOSSE) object tracker. Pyramidal convolutional features were then extracted for each tracked individual using a LiteFlowNet CNN. Finally, by learning the temporal differences in the frame sequences, a unique deep skip connection gated recurrent unit (DS-GRU) is trained to identify activities. Tested on public benchmark dataset, the suggested strategy produced state-of-the-art results, demonstrating its efficacy.

However, some hybrid approaches also incorporate sequential learning models like RNN [11] and LSTM [12] to enhance activity recognition accuracy. RNNs, particularly LSTMs and Bidirectional LSTMs, are powerful recurrent networks capable of preserving long-term dependencies. They have proven highly effective in applications such as speech recognition, human activity recognition (HAR), and gesture recognition [13]. Therefore, Bidirectional LSTM has been employed in this paper to recognize human actions. The flow diagram of the work is mentioned in the Fig. 1. This enables the proposed technique to efficiently learn long-term sequences. The key contributions of this work are outlined below:

- An important step toward activity recognition is fusion of multi-modalities. To achieve this, we utilize two different types of modalities such as RGB and depth video sequences.

- A lightweight 3DCNN is proposed as a spatial-temporal feature extractor for both RGB and depth sequences independently. The proposed lightweight 3DCNN is 20 times smaller and 1.26 times faster than the state-of-the-art CNN [14].
- Bidirectional LSTM(Bi-LSTM) is a kind of neural network to deal with time series data. To learn the spatial and temporal variations extracted in the previous step, we employed a Bi-LSTM network for both RGB and depth sequences. Subsequently, probability scores for each activity were obtained from both modalities.
- Finally, decision-level fusion was applied to aggregate the scores obtained from both RGB and depth data for the final recognition of human actions.
- The approach given in this work has been trained and tested on a benchmark UTKinectAction3D dataset and results show the efficacy of the method.

The rest of the paper is structured as follows: Section II provides a review of relevant literature. Section III presents the proposed methodology for activity recognition using RGB-D data. Section IV details the experimental work, results, and discussion. Section V compares the proposed approach with existing methods, section VI presents comparison with existing work and finally, Section VII concludes the paper and discusses future work.

2. Related Works

RGB-D based activity recognition is a prominent research area from several years. In this section, we have reviewed previous studies that utilized multi modalities such as RGBD.

The recent development in the human activity recognition in RGBD videos has gained significant attention from past few years due to the multimodality of data and robust deep learning algorithms for image and video classification task. Numerous comprehensive surveys are available in past literature to explore various human activity-based deep learning methods. In this work, some of the essential approaches based on lightweight CNN, RNN and fusion in RGBD have been discussed. For instance, Noor et al. [15] proposed a lightweight skeleton based 3DCNN for action recognition that outperforms significantly in terms of execution time and accuracy over baseline method. They have also suggested pose estimation during falling motion in the video sequences. Additionally, Lin at el. [16] proposed a lightweight neural network named LIMUNet and its variant LIMUNet-Tiny to recognize smartwatch based human activity recognition. The proposed network uses depth-wise separable convolutions and residual block in order to remove computation overhead and number of parameters. Validation on PAMAP2 and LIMU two benchmark datasets shows an improvement in accuracy by 2.9% using LIMUNet. Similarly, Gupta et al. [17] proposed a vision transformer and lightweight CNN based on end-to-end edge computing system and validated on three benchmark datasets: American Sign Language, NUS Hand Posture dataset and Turkey Ankara Ayrancı Anadolu High School's Sign Language Digits dataset and achieve an accuracy between 90-99% on test data. Moreover, to address the challenges in the lightweight network and HAR accuracy Huan et al. [18] recommended a lightweight hybrid Vision Transformer [LH-ViT]. The results obtained show the effectiveness of the recommended approach. In the same way, Choudhury et al. [19] introduced a hybrid, efficient, and lightweight CNN-LSTM model designed to extract both spatial and temporal features directly from raw sensor data in uncontrolled environments. The model classifies six daily living activities without relying on any data preprocessing or augmentation. Despite this, it outperformed all conventional models and achieved a remarkable 98% accuracy with optimized computational efficiency. Similarly, Lizo et al. [20] also applied lightweight CNN-LSTM model for dynamic hand gesture recognition which offers a powerful architecture for processing and making prediction based on the sequential data. They used lighter version of MobileNetV1 by removing one layer and got 98% recognition accuracy over four C, N, P and S alphabets. Verma et al. [21] discussed various machine learning based algorithms to handle the challenges activity recognition task. An optimized 3DCNN [22] is proposed to recognize human activities and obtains highest precision score which is 87.4%. Similarly, Verma et al. [23] also proposed 2DCNN to solve HAR problem and proposed algorithm was tested on UTKinectAction3D dataset to analyze the performance of the algorithm. Therefore, after above rigorous discussion the outstanding performance of lightweight 3DCNN with LSTM architecture motivative us to apply it in this proposed work.

Alternatively, recurrent neural networks mainly Bidirectional-Long Short-Term Memory Network (Bi-LSTM) is more suitable to deal with the HAR problem due to its ability to capture long term sequence with processing of inputs bidirectionally with high computation complexity and large number of parameters. To cater this problem, AlMuhaideb et al. [24], proposed a novel approach consisting of standard and residual LSTM with CNN. The method replicates the context awareness offered by Bi-LSTM employing data flipping augmentation as well. The results show the effectiveness of the suggested method (Res-LSTM) over traditional Bi-LSTM. It achieves 96.34% accuracy with 576,702 parameters when tested over UCI-HAR benchmark dataset. The method is also tested on KU-HAR datasets and achieves an accuracy of 97.20% with 192,238 parameters. Similarly, Basly at el. [25] suggested residual based CNN (RCN) to retain appearance and long-term dependencies among the sequences. The evaluation of the proposed approach was performed on two benchmark datasets such as CAD-60 and MSRDailyActivity3D and obtain very promising results when compare to the existing state-of-the-art. Transfer learning methods have also been applied to the activity recognition task [26], achieving state-of-the-art results.

Furthermore, various multi-modal based fusion approaches have been utilized for activity recognition task and achieved a great success in terms of accuracy of the actions. For instance, Verma at el. [27] proposed multi modal fusion based on RGB and Skeleton video sequences. They extracted MHI and MEI information from RGB sequences and generated skeleton images from skeleton joints. Later on, they applied a decision level fusion approach and fused the score obtained from all parallel streams corresponding to RGB and Skeleton. This fusion approach was validated on two benchmark datasets namely: UTD-MHAD and NTU-RGB+D120 and achieved reasonable results compared to the existing state-of-the-art methods. Similarly, Shafizadegan et al. [28] summarized various multi-modal based fusion strategies along with publicly available HAR datasets, comparisons and future work directions. Likewise, Rehman et al. [29] also presented a multi-modal based analysis comprising of skeleton tracking, RGB imaging and pose estimations. The proposed work eliminates the drawback of using unimodality by leveraging the advantages of both RGB and skeleton modalities. The proposed work was trained and tested on the benchmark multi-modal HAR dataset namely: UTD-MHAD. The finding of the work illustration its robustness, reliability and ability to apply in real time of HAR system. Therefore, based on the above discussion we have applied lightweight 3DCNN and Bi-LSTM model to handle the problem of activity recognition in RGBD video sequences.

3. Proposed Methodology

In this section, the proposed framework to recognize human actions and their main components are discussed in details containing an action A_I in the sequence of frames of RGB-D video V_I using proposed lightweight 3DCNN-BiLSTM model where 3DCNN is used for spatial-temporal feature extraction for F_N frames from two different modalities RGB and depth simultaneously. The proposed work is divided into three parts: Firstly, spatial-temporal features are fetched from both RGB and depth videos V_I with a jump J_F frames such that the skip of J_F does not disturb the dynamic information available in the action A_I in the video sequence. Second, the extracted spatial-temporal features representing the action A_I are fed into Bi-LSTM in C_N chunks, where C_N represents the features of a video frame. At the last, the probability scores are generated and fused using decision level score fusion approach (DLSF).

3.1. Preprocessing and Spatial-temporal Feature Extraction using 3DCNN from Video Sequence

A. Resizing

The RGB-D video sequences collected from the UTKinectAction3D dataset have original resolutions of 480x640. The proposed method processes these RGB-D video sequences by resizing them to a new dimension of 32x32 using the OpenCV library in Python. In the original video sequence, each frame contains 3 RGB channels, which are converted into grayscale images with a single channel. Since the proposed approach operates on a fixed number of frames, and each RGB-D sequence contains a varying number of frames so, each sequence has been resized to a fixed length which is 20 for both datasets before being input into the 3DCNN for spatial-temporal feature extraction.

B. Spatial-temporal Feature Extraction

Convolutions are carried out on video sequences in order to extract features 2D feature maps based only on the spatial dimensions. However, during the video analysis, it is essential to record the motion information embedded in several successive video frames. To address this, we propose 3D Convolutional Neural Network (3DCNN), introduced in [30], is a deep learning model built to extract features from both spatial and temporal dimensions simultaneously. It achieves 3D convolution by applying a 3-dimensional filter across a cube formed by stacking spatial and temporal frames sequentially. To capture motion-related information from frame sequences, the feature maps within the convolutional layers are linked to multiple consecutive frames from the preceding layer. Multiple convolutional layers are utilized to learn both low-level and high-level features. The design approach for convolutional networks includes adding more layers to enhance the feature maps. Consequently, a 3D CNN is constructed by convolving a 3D kernel over stacked frames, generating a 3D cube. The value at the (x, y, z) position in the j^{th} feature map of the i^{th} layer is defined by equation (1).

$$v_{i,j}^{x,y,z} = \tanh\left(b_{ij} + \sum_{m}\sum_{a=0}^{A_{i-1}}\sum_{b=0}^{B_{i-1}}\sum_{c=0}^{C_{i-1}} w_{ijm}^{abc} v_{(i-1)m}^{(x+a)(y+b)(z+c)}\right) \quad (1)$$

Where $v_{i,j}^{x,y,z}$ represent the output feature map value at position (x, y, z), m is the index iterating over the feature maps from the previous layer i-1, b_{ij} is the bias term associated with j^{th} feature map of the i^{th} layer, w_{ijm}^{abc} is the convolutional kernel weights connecting the m^{th} feature map of layer i-1to the j^{th} feature map in layer i. The variable a, b, c represent the specific position of 3D kernel. $v_{(i-1)m}^{(x+a)(y+b)(z+c)}$ is the value at position $(x+a, y+b, z+c)$ in the m^{th} feature map from the previous layer (i-1). The next section introduces two different lightweight 3DCNN feature extractors have been proposed for both RGB and depth video inputs.

- Spatial Temporal feature extraction using Lightweight 3DCNN from RGB video sequences

In this work, 3DCNN serves as a feature extractor that utilizes three-dimensional input data, offering better adaptability to the continuous temporal and spatial domain characteristics inherent in video sequences. The proposed lightweight spatial-temporal 3DCNN feature extractor is implemented using a sequential architecture. To prepare the input for the lightweight 3DCNN model, all video sequences were standardized to a fixed depth of 20 frames, with each frame resized to 32x32 pixels, resulting in an input vector of size (20x32x32) where the size of both length and width is 32 as discussed in the pre-processing section. This proposed 3DCNN architecture consists of only two convolutional layers and two max-pooling layers, which is why it is referred to as a lightweight architecture. The architecture starts with a 3D convolutional layer (C1) that includes 32 filters with a kernel size of (3×3×3), utilizing ReLU activation to introduce nonlinearity and L2 regularization (0.001) to mitigate overfitting. The input shape for this layer is set to (20,32,32,1). Following the C1 layer, a dropout layer with a rate of 0.2 is applied, succeeded by a max-pooling layer (M1) with a pool size of (2×2×2). Next, a second 3D convolutional layer (C2) with 64 filters and a kernel size of (3×3×3) is employed, followed by ReLU activation. After C2 layer, another dropout layer with a rate of 0.3 is applied, followed by a second max-pooling layer (M2) with a pool size of (2×2×2). The output is then passed through a time-distributed flattening layer, preserving the temporal structure of the feature maps for further processing by a bidirectional long short-term memory (Bi-LSTM) network. The proposed architecture is designed to maintain computational efficiency while effectively capturing both spatial and temporal features. Fig. 2 illustrates the 3D CNN architecture used for spatio-temporal feature extraction from RGB video sequences.

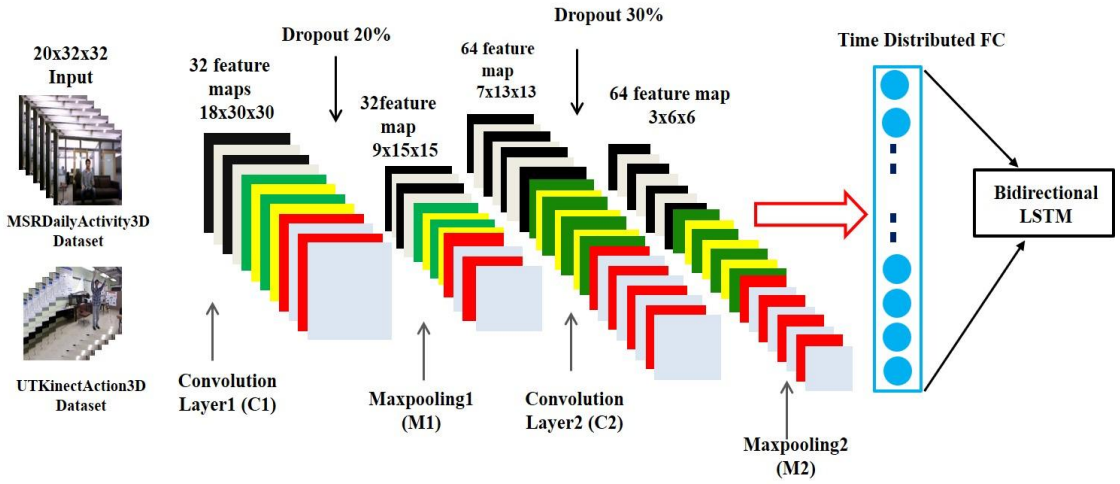

Fig.2. Architecture of spatial temporal 3DCNN feature extractor for RGB video sequences

- Spatial Temporal feature extraction using Lightweight 3DCNN from Depth video sequences

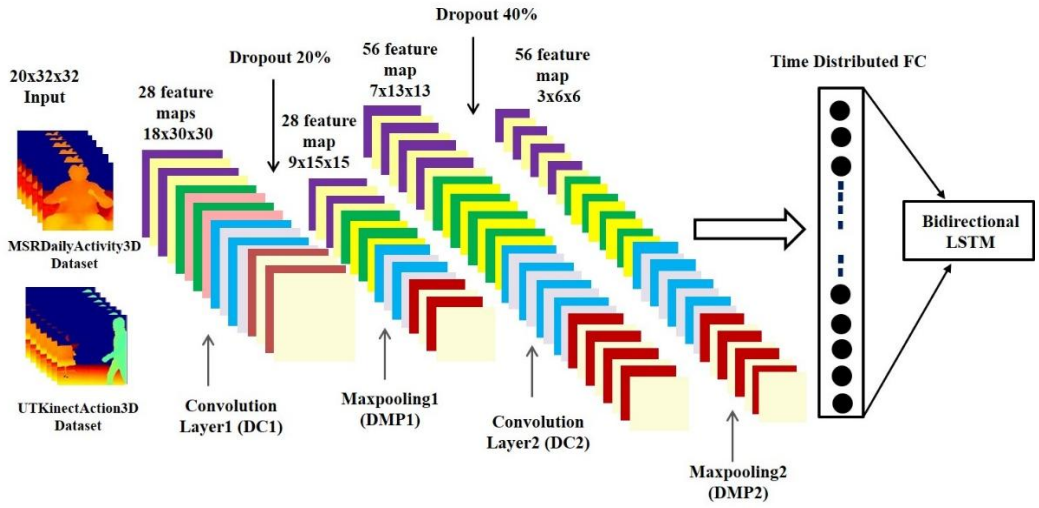

Fig.3. Architecture of spatial temporal 3DCNN feature extractor for Depth video sequences

In order to efficiently process depth video sequences, another lightweight spatio-temporal 3DCNN feature extractor is proposed. The architecture begins with a 3D convolutional layer (DC1) containing 28 filters having a kernel size of (3×3×3), using ReLU activation. The input shape of DC1 layer is set to (20,32,32,1), representing a sequence of 20 grayscale frames with dimensions of 32×32. After the DC1 layer, a dropout layer with a rate of 0.2 is applied to mitigate

overfitting. This dropout layer is followed by the first max-pooling layer (DM1) with a pool size of (2×2×2) to perform spatial and temporal down-sampling. Next, a second 3D convolutional layer (DC2) is introduced, consisting of 56 filters with a kernel size of (3×3×3) and utilizing ReLU activation. This layer is followed by a dropout layer with a rate of 0.4 followed by a second max-pooling (DM2) layer with the same pool size as used in the previous max-pooing layer (DM1). Finally, the output is sent to the time-distributed flatting layer, keeping the temporal structure preserve in the feature maps. The proposed architecture balances the analysis performance by maintaining the spatial-temporal features that will be fed into the bidirectional LSTM (Bi-LSTM) model. The architecture of the 3DCNN used for spatio-temporal feature extraction from depth video sequences is illustrated in Fig. 3.

C. Activity Score Prediction using Bi-LSTM with 3DCNN Features

RNNs are mostly used for inspecting the hidden sequential patterns available in both spatial sequential and temporal sequential data [31]. Another kind of sequential data, where the movement of the visual content are illustrated through a series of frames such that these frame sequences assist to understand the context of the action. Such sequences can be interpreted by RNNs, however in case of the long-term sequences, they forget the earlier input of the sequence. This problem is referred to as vanishing gradient problem and can be resolved using a specific type of RNN called LSTM [32]. It is specifically designed to capture the long-term dependencies, where it controls long term pattern sequence is controlled by input, output and forget gates.

- 2-layers Bidirectional LSTM (Bi-LSTM)

One type of recurrent neural network (RNN) that may identify long-term dependencies in sequential data is LSTM. In this work, a variant of LSTM known as Bi-directional Long Short-Term Memory (BiLSTM) is used, which leverages information from both past and future contexts within the sequence [33]. This method works particularly well for video action recognition, since the relationships between actions depends on the scene before and after. The LSTM cell utilized in our system is characterized in equations from (2) to (7).

$$i_t = \sigma(b_i + U_i x_t + W_i h_{t-1}) \tag{2}$$

$$f_t = \sigma(b_f + U_i x_t + W_f h_{t-1}) \tag{3}$$

$$o_t = \sigma(b_o + U_o x_t + W_o h_{t-1}) \tag{4}$$

$$g_t = \sigma(b_g + U_g x_t + W_g h_{t-1}) \tag{5}$$

$$c_t = f_t c_{t-1} + g_t i_t \tag{6}$$

$$h_t = \tanh(c_t) o_t \tag{7}$$

Where σ represents the sigmoidal function, W_i W_f W_o represents the weight matrices for the input, forget and output gates, respectively. Similarly, b_i , b_f, and b_o are the corresponding bias vectors for input gate, forget gate and output gate. The i_t is the input gate which controls the new informaiton flow in the cell, whereas the forget gate f_t determinds which information should be discarded from the cell's internal state. The o_t output gate control the information that is passed to the output. The input modulation gate g_t servers as a primary input to the cell. c_t is the internal state that manages internal recurrence of the cell and h_t is the hiddent state that retains the previous infromation. The structure of LSTM cell structure is shown in the Fig.4.

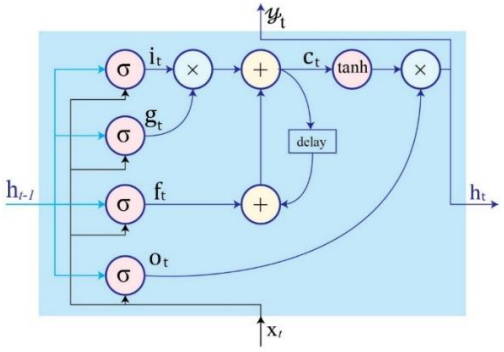

Fig.4. Diagram of an LSTM cell structure showcasing its internal recurrence mechanism [34]

The selection of the number of layers in a BiLSTM network is a pivotal factor in designing the architecture for an action recognition task. Determining the optimal number of layers depends on various parameters such as complexity of the recognition task, computational resources available and dataset size. Using many layers in BiLSTM network can enhance its capacity to learn intricate relationship exist within the data. However, increasing the number of layers also raises the potential risk of overfitting, particularly when working with smaller datasets.

A typical strategy is to start with a single layer and gradually increase the number of layers as needed. Maintaining a balance between the number of layers and the total parameters in the model is crucial to reducing the risk of overfitting. In our approach, we chose two Bi-LSTM layers, considering the moderate size of our datasets and the complexity of the action recognition task. The capacity of the model and the risk of overfitting are effectively balanced by this configuration. Furthermore, the network may extract more abstract information from the input when it has multiple layers, which improves the model's accuracy.

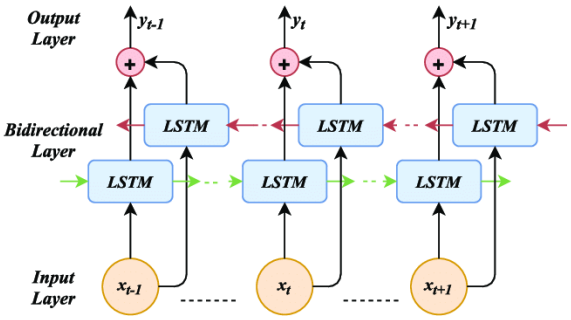

Fig.5. Structure of bidirectional LSTM layer

Another crucial aspect of designing a BiLSTM network's architecture is selecting the appropriate number of nodes for each layer. The number of nodes in each layer determines its capacity to extract information from the input. While a larger number of nodes in the network enhances its capacity and it also increases the count of the parameters and the likelihood of overfitting. For this implementation work, we selected 64 nodes per BiLSTM layer. The shape of BiLSTM layer is shown in the Fig.5 where the bidirectional layer is a layer which contains both forward and backward pass.

The number of nodes was chosen based on the complexity of the action recognition task, the available computational resources, and the size of the dataset. The space time features extracted by 3DCNN in previous phase are used to train 2-layes BiLSTM network. After second BiLSTM layer a dense layer of 64 neuron is used followed by a dropout layer with an amount of 20%. This dropout layer is followed by the SoftMax layer for output classification. During the network training rectified linear unit 'ReLU' activation function is applied in both BiLSTM layers.

Fig. 6 depicts the 2-layer BiLSTM network designed for modeling activity sequences using the spatial-temporal 3DCNN features extracted earlier. It also highlights that both the first and second layers of the BiLSTM network utilize 64 units each.

In this work, we have chosen Bi-LSTM over GRU because of its superior capacity to capture temporal dependencies in both forward and backward directions, especially beneficial for action recognition task where understanding of entire sequence context is essential.

According to Chung et al. [35], GRUs are computationally more efficient and require fewer parameters, but they are generally less expressive than Bi-LSTMs when it comes to capturing complicated and symmetric motion patterns. It has been demonstrated that Bi-LSTMs enhance performance in sequence modeling tasks including voice recognition [36], gesture recognition [37], and video-based action classification [38] by processing sequences in both directions.

Transformer-based or attention models provide state-of-the-art performance, but they are less appropriate for the UTKinect-Action3D dataset utilized in our work since they often demand larger datasets and more processing power [39]. BiLSTM achieves a proper balance between computational complexity and temporal modeling.

D. Decision Level Score Fusion (DLSF)

The decision-level fusion (DLSF) method has been employed to predict probability scores for every RGB and depth video test sequence. This approach processes classification outputs from both RGB and depth streams and combines them to produce more reliable and accurate results. There are two primary methods for implementing DLSF: supervised and unsupervised. The supervised approach requires additional training, where the scores, outputs, or labels from individual classifiers serve as input to a training model. Conversely, the unsupervised approach does not involve any extra training. Popular unsupervised methods for integrating individual classifiers include the Weighted Product Model (WPM), Average Method, Borda Count, and Majority Voting. In this study, the WPM technique was adopted as the DLSF method.

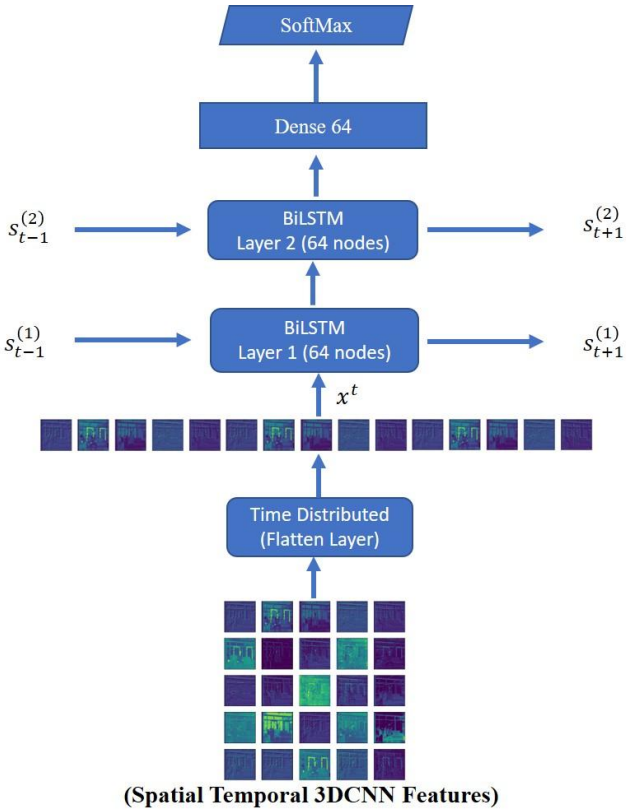

Fig.6. 2-layer Bi-LSTM network for modeling activity sequences using the spatial-temporal 3DCNN features

E. Weighted Product Model (WPM)

The Weighted Product Model (WPM) is a widely recognized decision-level fusion technique depended on the principles of multi-criteria decision-making (MCDM). In MCDM, the key challenge lies in integrating the probability scores obtained by RGB and depth video streams. The main distinction between WPM and the Weighted Sum Model (WSM), despite their basic similarity, is that WPM employs multiplication rather than addition. In this method, the predicted probability scores are combined to identify the most suitable class for a given activity sequence. This paragraph provides a concise explanation of the Weighted Product Model (WPM) method. In a given MCDM problem, assume there are P decision statements and Q alternatives. Let r_j represent the relative weight of the j^{th} statement, and S_j denotes the performance value of alternative A_k corresponding to statement j. The WPM can then be mathematically expressed as mentioned in the equation (8):

$$P(A_k) = \prod_{j=1}^{n} (a_{k_j})^{w_j} \quad \text{for } k = 1, 2, 3\ldots \text{ m.} \tag{8}$$

Let S_{RGB} and S_{depth} represents the probabilities score generated for each test sequences of RGB and depth videos streams. These scores act as decision-making criteria however, the number of actions is determined by the alternatives. Each stream produces a score matrix vector with equal number of actions. The most suitable activity, denoted by the highest value of (A_k), is chosen based on this structure. Let W_1 and W_2 are the weights assigned to each stream and value 1 is assigned for each dataset. When all the weights are equal, the Weighted Product Model (WPM) leads to product rule. Equation (9) mathematically expresses the fusion of scores using WPM method. In this work, we evaluated early, intermediate, and decision-level fusion strategies on the UTKinect-Action3D dataset. Decision-level fusion (WPM) achieved the highest accuracy of 96.72%, compared to 93.68% for early fusion and 92.59% for intermediate fusion, as shown in Table 1.

$$WPM_s = \max[(S_{RGB}^{w_1}) \times (S_{Depth}^{w_2})] \tag{9}$$

4. Experimental Results and Discussions

In this study, UTKinectAction3D [40] public RGBD dataset is utilized to validate the given approach. This is a RGBD dataset because this was captured using Microsoft Kinect Senor. The dataset's activities are matched across all formats, including RGB, Depth, and Skeleton. For the training of our model, we used 12th Gen Intel(R) Core (TM) i7-

1255U 1.70 GHz with 16GB of RAM running on Ubuntu 22.04 LTX version. Further, to implement 3DCNN and LSTM network we used Google Collaboratory.

4.1. UTKinectAction3D Dataset Description

The UTKinectAction3D dataset was generated by Microsoft Foundation Research (MSR) in 2012. The dataset comprises 10 actions performed by 10 persons, including nine male and one female. The recorded actions are 'carry', 'walk', 'sit down', 'stand up', 'push', 'pull', 'throw', 'pickup', 'wave hands', and 'clap hands. Every person repeats the action two times and the data is recorded in three different formats such as RGB, depth and skeleton. An action is synchronizing in all three formats. One sequence of carry action is missing in the dataset, resulting 10(users)*10(actions)*2(repetition) = 200, but in reality, it was 199 because one carry sequence is missing. The total action sequences in the dataset was 200 * 3 = 600 pertaining to the RGB, depth and skeleton data. RGB, depth and skeleton data exist in the dataset in the form of jpg, bin and txt files respectively. In MSR laboratory the dataset was gathered with the help of stationary Kinect camera and all ten actions have been shown in the Fig.7.

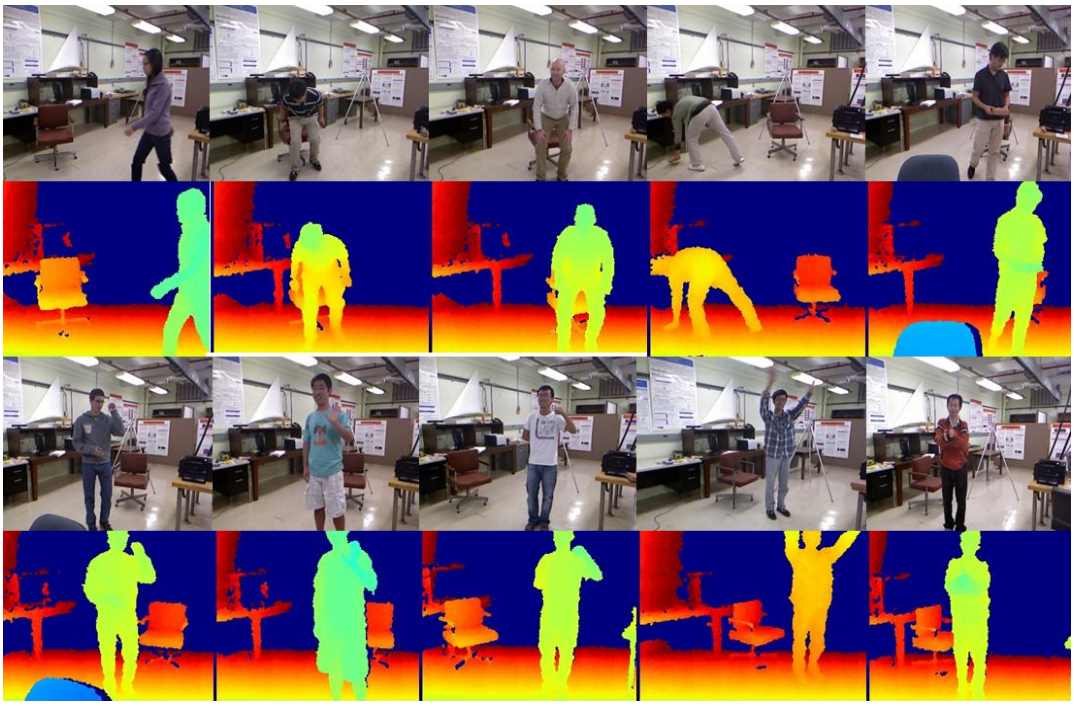

Fig.7. Actions from the UTKinectAction3D dataset in RGD and depth format, namely walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands, from top to bottom and left to right, respectively

4.2. Experimental Results on UTKinectAction3D Dataset

The primary goal of our proposed methodology is to recognize the human activities from 3D data such as videos. For this purpose, our proposed approach utilizes two modalities such as RGB and depth maps. The suggested approach operates in three different stages such as space-time feature extraction through Lightweight 3DCNN from RGB and depth video sequences, temporal sequence learning with the help of BiLSTM and activity score fusion using weighted product method. During the experimentation varying size of input like (18×32×32), (19×32×32), (20×32×32), (21×32×32) and (22×32×32) were applied but obtained the best features corresponding to the input with size (20×32×32) for both RGB and depth sequences. During model training the learning rate of .0001 was with a decay parameter which gradually decreases with increase of the epoch counts. The model was trained using the Adam optimizer, with categorical cross-entropy as the loss function. The UTKinectAction3D dataset is used to validate the proposed approach and the visualization of their samples have been mentioned in the Fig.7.

To validate our proposed approach, we employed leave-one-out cross-validation (LOOCV). In LOOCV, various training and test set splits are used, such as 90:10 (9 users for training and 1 for testing), 80:20 (8 users for training and 2 users for testing), 70:30 (7 users for training and 3 for testing) and 60:40 (6 users for training and 4 for testing), along with their corresponding results. For instance, in case of 90:10 split, 9×16×2=288 sequences are utilized for training purpose and 1×16×2=32 sequences are utilized for testing purposes. The four different types of train and test split have been shown in the Fig.8. Furthermore, the extracted space-time features are utilized by BiLSTM network for sequence learning, and the scores are recorded for each activity sequence for further processing.

Fig.8. Training and test set splits

The visualization of feature maps for ‘wave hands’ activity of UTKinectionAction3D dataset after first layer of our proposed Lightweight 3DCNN feature extractor for RGB sequences is shown in the Fig.9. The ROC curve of the UTKinectAction3D dataset is also shown in the figure in the Fig.10. The performance of our model at various threshold levels is represented by the curve that is drawn between the true positive rate and the false positive rate.

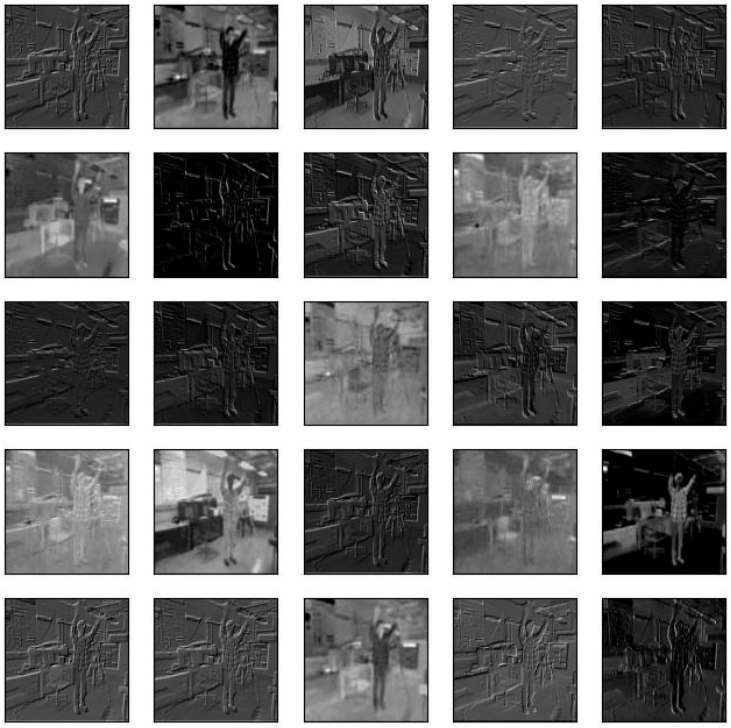

Fig.9. Visualization of feature maps after first convolution layer of ‘wave hands’ activity for UTKinectAction3D dataset

The confusion matrix of UTKinectionAction3D dataset is shown in the Fig.11 and illustrate that the model achieves high accuracy, and properly classifying most of the actions, however some misclassification also exists due to similarity between the actions. The highest confusion occurs between "push" and "pull" actions with "push" being incorrectly classified as "pull" (22.7%), due to the high level of similarities between the hand motions. It also reflects the inherent similarity between these actions. Moreover, both actions involve forward or backward arm movements along the same plane, making their motion trajectory visually and kinematically similar. Further, the inter-subject variability and dataset's viewpoint constraints in how push and pull are performed generate additional overlap, and supporting to more confusion. These factors reduce the discriminative features available to the model, justifying the observed confusion in classification. Similarly, "carry" action is misclassified with actions "walk"(2.3%) and "clapHands"(0.8%) due to the overlap in movement pattern. Likewise, "SitDown" and "Standup" actions are also misclassified (4.1%), just because of similar posture between them. The high classification accuracy of "waveHands" (99.5%) indicates that the model successfully

separate it from other actions. The "Pickup" action achieves 100% classification accuracy suggests that it contain distinct features. However, sometimes "throw" action is misclassified as "walk"(3.9%) action due to the similarity in the arm motion. Minor misclassification of "capHands" with "carry" (0.8%) suggests slight overlap in hand movements. Overall, the model performs well with 96.57% accuracy, but minor errors mainly occur in actions with similar spatial and temporal patterns.

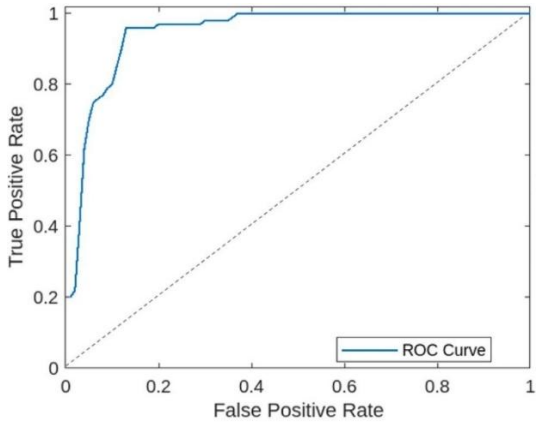

Fig.10. ROC curve of UTKinectAction3D dataset

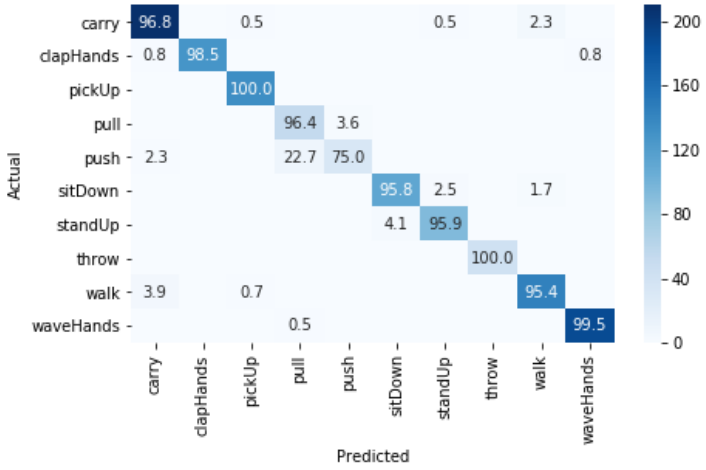

Fig.11. Confusion Matrix of our proposed approach for UTKinectAction3D dataset

Additionally, the training and accuracy curve of UTKinectAction3D dataset is also shown in the Fig.12 and Fig.13 respectively. It is evident from the figure 11 that model training runs for almost 300 epoch and at epoch number 265, validation loss is minimum 0.1706 and model attain highest accuracy of 96.72% at the same epoch in accuracy curve shown in the figure 12.

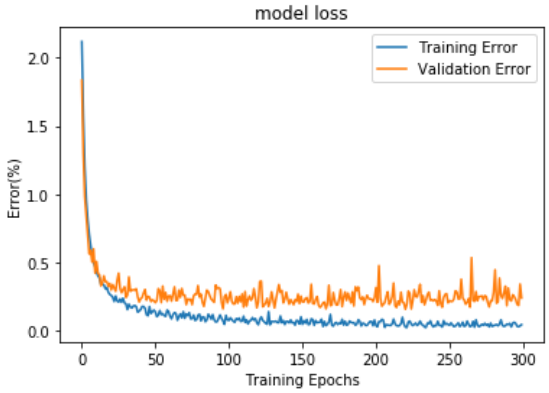

Fig.12. Training curve of UTKinectAction3D dataset

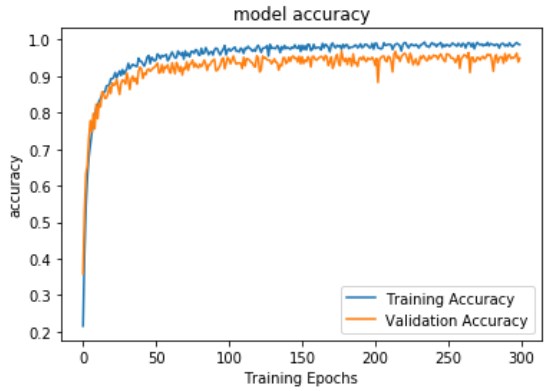

Fig.13. Accuracy curve of UTKinectAction3D Dataset

Table 1. No of parameters used in lightweight 3DCNN feature extractor

Layer (type)	Output Shape	Parameters
Conv3d (Conv3D)	(None, 18, 30, 30, 28)	784
dropout (Dropout)	(None, 18, 30, 30, 28)	0
max_pooling3d(MaxPooling3D)	(None, 9, 15, 15, 28)	0
Conv3d_1(Conv3D)	(None, 7, 13, 13, 56)	42392
dropout_1(Dropout)	(None, 7, 13, 13, 56)	0
max_pooling3d_1(MaxPooling3D)	(None, 3, 6, 6, 56)	0
time_distributed (Time Distributed)	(None, 3, 2016)	0
bidirectional (Bidirectional)	(None, 3, 128)	1,065,472
bidirectional_1 (Bidirectional)	(None, 128)	98,816
dense (Dense)	(None, 64)	8,256
dropout_2 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 16)	1,040
	Total Parameters	**12,16,760 (4.68 MB)**

During feature extraction from both RGB and depth sequences, the total number of parameters are 1,216,760 (4.68 MB), with all of them being trainable and none non-trainable. As evident from Table 1, there are no non-trainable parameters since every parameter is trainable. The proposed lightweight 3DCNN network efficiently utilizes only 12.16 million trainable parameters to capture both spatial and temporal. Our proposed 3D CNN model is significantly more lightweight compared to several widely used 3D CNN architectures. For instance, C3D [41] and ResNeXt-101 (3D) [42] contain approximately 78M and 83M parameters, respectively, while I3D [43] has 25M parameters. In contrast, our model contains only 12.71M parameters, making it over 84.41% smaller than C3D and 85.34% smaller than ResNeXt-101(3D) model. This substantial reduction in parameter count underscores the computational efficiency and enhances deployablity of our model, especially for real-time or resource-constrained applications.

5. Comparative Results Analysis

In order to demonstrate the effectiveness of the suggested method, we evaluated its performance against that of many classifiers, including KNN-based action recognition, SVM-based action recognition, and 3DCNN-based action recognition. In this work, a novel approach Lightweight 3DCNN + Bi-LSTM model for action recognition has been proposed. Accuracy, precision and recall and have been used to analyze the performance of the proposed approach.

Fig.14 analyzes the accuracy of the proposed recommendation system by adjusting the training and test sets size. In the recommended approach, two lightweight 3DCNN networks are used during the recognition process. Both lightweight 3DCNN are used as feature extractors from RGB and Depth video sequences. Next, various classifiers, including KNN, SVM, 3D-CNN, and Bi-LSTM, are trained using the features extracted in the previous step. The lightweight hybrid approach used in the proposed work increases the recognition accuracy. In order to validate the effectiveness of the presented work, we compare our proposed method with different trained classifiers such that K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and 3DCNN itself. According the Fig.14, the proposed method attains a maximum accuracy of 96.72%, for 80:20 dataset split, which is 84.79% for KNN based action recognition, 78.39% for SVM based action recognition and 83.25% for 3DCNN based action recognition itself. This is due to the hybrid lightweight 3DCNN-BiLSTM and two-pronged fusion process.

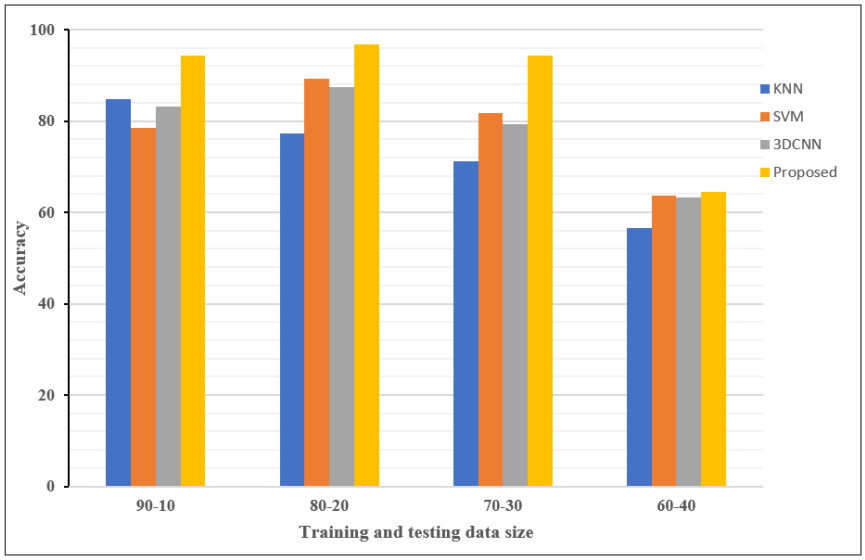

Fig.14. Analysis of performance based on accuracy

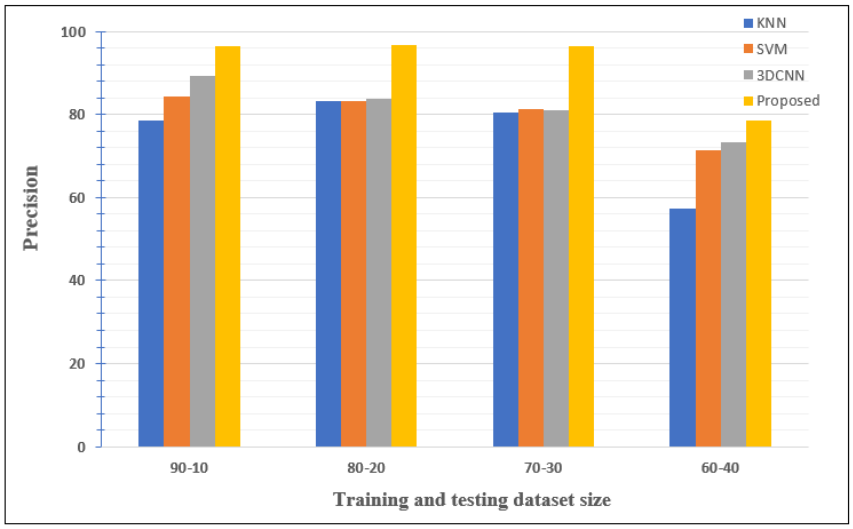

Fig.15. Performance of the proposed approach based on precision

In Fig.15 evaluate the performance of the suggested method based on precision. As shown in Fig.15, the proposed method achieves precision values of 96.32%, 96.68%, and 96.38% for dataset splits of 90:10, 80:20, and 70:30, respectively as compare to the KNN, SVM and 3DCNN.

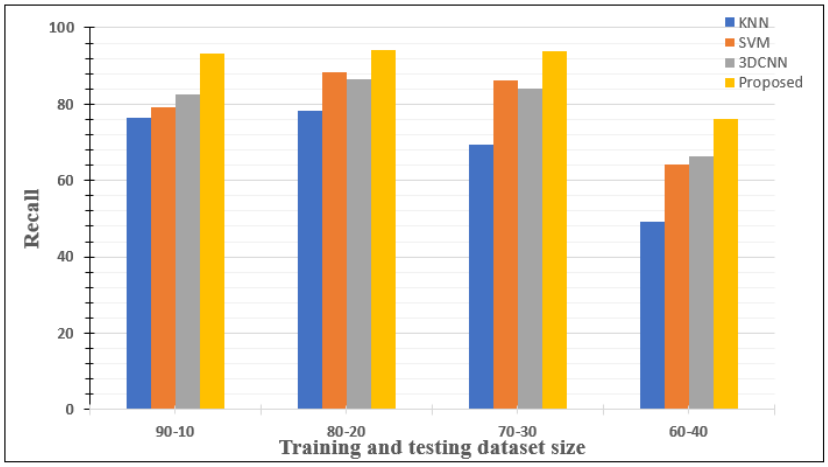

Fig.16. Performance of the proposed approach based on recall

Again, Fig.16 evaluates the performance of the suggested method based on recall value. It is clear from the Fig.16 that a good dataset split has highest recall value which is 94.03% for 80:20.

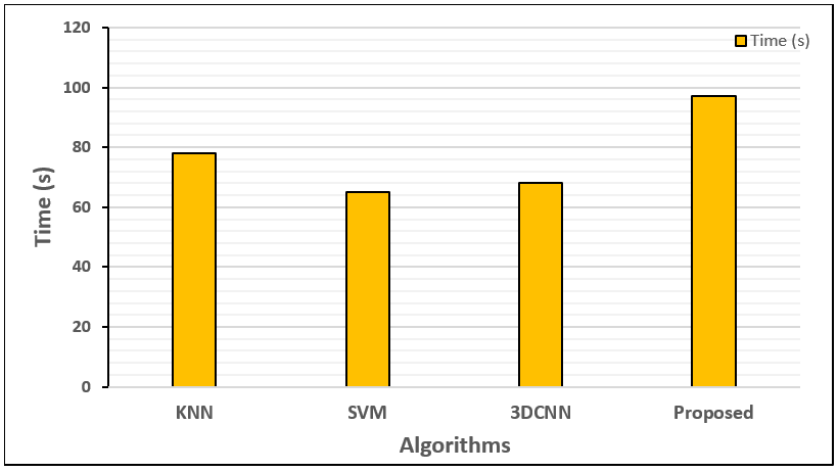

Fig.17. Analysis of complexity of various algorithm based on time

In Fig.17, the time complexity of various algorithms has been analyzed. The time complexity of an algorithm quantifies the amount of time it requires to execute. After analyzing Fig.17, it is clear that the proposed approach takes maximum time to execute as compared to the other algorithms. This is due to the hybrid approach used in this work. The processing time has no impact on the overall recognition accuracy.

Our model is designed to be lightweight and less computationally complex in term of parameters and FLOPs, with approximately 1.2 million parameters—significantly fewer than standard 3DCNNs such as C3D (~78M) and I3D (~25M). The computational complexity of proposed model is 35.4 MFLOPs (million FLOPs), which is significantly lower than C3D(~38 GFLOP) and I3D (~107 GFLOP). Hence, the proposed architecture has a lower number of FLOPs, making it more suitable for real-time deployment.

6. Ablation Study

We carried out a detailed ablation study with UTKinectAction3D dataset to assess the contribution of each component in our proposed action recognition framework. This analysis meticulously examines the effect of each of the main elements such as: the fusion method (early, intermediate, and decision), the temporal analysis (e.g. Bi-LSTM, GRU), and the modality (RGB, depth and both). This allow for better understanding of how each component improves the model's functionality and support the final architecture.

An extensive ablation study was conducted on the UTkinectAction3D dataset in order to evaluate the contribution of each element within the suggested 3DCNN-BiLSTM framework. The study first starts to assess the performance of two distinct modalities, RGB and Depth, processed independently using lightweight 3DCNN-BiLSTM model. During independent modality processing, RGB sequences achieved an accuracy of 89.32%, while the depth sequences performed slightly better at 90.75%, showing that the depth modality is more robust during capturing structure motion patterns. Compare to using simply the 3DCNN, the BiLSTM component greatly improve the temporal learning capacity, demonstrating the importance of temporal dynamics in recognizing complex human actions [38, 41].

Further, an analysis of various fusion strategies was conducted. Early fusion (feature-level), intermediate fusion, and decision-level score fusion (DLSF) using Weighted Product Model (WPM) were the three fusion types that were evaluated. The decision-level fusion technique performed the best among them, achieving the greatest accuracy of 96.72%, while early fusion and intermediate fusion achieved 93.68% and 92.59%, respectively. This shows that compared to joint feature aggregation, independent stream modelling followed by decision fusion offers stronger discriminative insights into spatiotemporal patterns.

Additionally, for temporal modelling, the study compared BiLSTM with GRU models. BiLSTM was selected because of its bidirectional context modelling, which produced a better accuracy (96.72%) than GRU (94.13%), which is consistent with results from earlier gesture and action recognition studies [35,36]. Table 2 shows the summarization of the findings discussed above.

All these findings show how important each module is to improving the overall performance of the system, especially the lightweight dual-stream 3DCNN and the BiLSTM-based temporal modelling.

Table 2. The table below summarizes these findings, providing a clear comparative perspective on how different architectural choices impact recognition performance

S No.	Model Configuration	Modalities	Temporal Model	Fusion Method	Accuracy (%)	References
1	3DCNN only (no temporal model)	RGB	None	None	81.57	[41]
2	3DCNN+BiLSTM	RGB	BiLSTM	None	89.32	[38]
3	3DCNN+BiLSTM	Depth	BiLSTM	None	90.75	Proposed
4	3DCNN+GRU	RGB + Depth	GRU	Decision-Level (WPM)	94.13	[35]
5	3DCNN+BiLSTM	RGB + Depth	BiLSTM	Early Fusion	93.68	[36]
6	3DCNN+BiLSTM	RGB + Depth	BiLSTM	Intermediate Fusion	92.59	Proposed
7	**3DCNN+BiLSTM (Lightweight)**	**RGB + Depth**	**BiLSTM**	**Decision-Level (WPM)**	**96.72**	**Proposed (Final)**

7. Comparison with Existing Literature

To highlight the effectiveness of the proposed method, we compare our algorithm with several existing approaches, as presented in the Table 3. In this section, we evaluate our work with existing approaches available in the literature, including Wang et el. [7], Anirudh et al. [8], Liu et al. [9], Zhang et al. [10], Liu et al. [11] and Verma et al. [12]. In Table 2, the accuracy of the proposed work is 96.72%. In [7], a key-pose-motifs are used to carried out human action recognition. Here, UTKinectionAction3D dataset is utilized for the experimentation purpose. Here, a sequence is matched with the motifs of each class to classify an action, and the class with the highest matching score is chosen. In [8], a variant of PCA namely a low dimensional embedding with a manifold (mfPCA) function has been proposed to carried out the action recognition task. In [9], a multi model feature fusion strategy with gated LSTM module is proposed to perform the action recognition task on skeleton sequences. The proposed method effectively preserves long-term context representation within the memory cell. The proposed method utilizes both spatial and temporal features for action recognition task. In [10], a simple universal spatial modeling approach has been suggested that was perpendicular to the RNN network enhancement. For datasets are used to validate the proposed approach. The proposed method achieves state-of-the-art results, demonstrating its effectiveness. In [11], a time-invariant and view point invariant method consisting of SVM and space-time feature learning using 3D-based deep CNN ($3D^2$CNN) and Joint Vector fusion for action recognition. The proposed method used UTKinectAction3D and MSR-Action3D dataset are used for validation. The suggested method achieves comparable state of the art methods. In [12], Verma et al. proposed a multi-model approach using 3DCNN and SVM, later on they optimized the probability score using two evolutionary algorithms namely particle swarm optimization and Genetic Algorithm. The proposed approach is validated on two RGBD datasets such as MSRDailyActiviy3D and UTKinectAction3D. The method proposed in [12] achieves state-of-the-art results comparable to existing methods, demonstrating the efficiency of the proposed approach. The obtained results also demonstrate that Genetic algorithm outperform over PSO.

Table 3. Comparison with UTKinectAction3D dataset based on accuracy

Methods	Accuracy (%)
Wang et al. (2016) [44]	93.5
Anirudh et al., (2015) [45]	94.9
Liu et al., (2017) [46]	95
Zhang et al., (2017) [47]	95.9
Liu et al., (2016) [48]	96.00
Verma et al., (2021) [49]	96.50
Proposed Approach (RGB + Depth) early fusion	**92.59**
Proposed Approach (RGB + Depth) intermediate fusion	**93.68**
Proposed Approach (RGB + Depth) decision level fusion	**96.72**

8. Conclusions and Future Work

In conclusion, this study demonstrates the effectiveness of our presented approach for RGBD based action recognition, which employs lightweight 3DCNN based feature extractor and Bidirectional LSTM model. The finding from experiment on UTKinectAction3D benchmark dataset validate the effectiveness of the proposed method. In particular, our proposed method exhibits outstanding performance on the UTKinectAction3D benchmark dataset over existing methods. The achieved results are due to the utilization of a lightweight 3DCNN and Bidirectional LSTM architecture, which effectively capture the complex spatial and temporal structures present in RGBD videos for action

recognition.

Furthermore, the employment of weighted product model as a decision level fusion method has proven its ability to capture intricate spatiotemporal aspects, this reduces computational cost and improves the overall effectiveness of the presented method. The proposed method uses two channels and is able to execute on multi-core CPU system in order to save the time. Additionally, our approach has proven to be quite flexible, retaining good accuracy for a wide variety of movements, from easy to difficult. This versatility emphasizes its importance beyond the academic study by highlighting its wider use in various real-world applications such human-computer interaction, content-based video retrieval and surveillance.

Despite the several advantages, our proposed method also suffers from some limitations such as clutter background, recognition of multiple activities and multi person activities simultaneously. Future research would be focused on addressing these limitations by incorporation additional skeleton joint information, addition of attention mechanism. Various diverse benchmark datasets may be used to validate the method with additional evolution parameters.

References

[1] Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40(6), 1510–1517 (2017)

[2] Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Thirty-First AAAI Conference on Artificial Intelligence, pp. 4263–4270 (2017)

[3] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 2010, pp. 9-14.

[4] L. Xia, C.C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 2012, pp. 20-27.

[5] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 2013, pp. 486-491.

[6] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision,2013, pp. 3551–3558.

[7] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, J. Huang, End-to-end learning of motion representation for video understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.6016–6025.

[8] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[9] Y. Shi, Y. Tian, Y. Wang, T. Huang, Sequential deep trajectory descriptor for action recognition with three-stream CNN, IEEE Trans. Multimed. 19 (7) (2017) 1510–1520.

[10] Ullah, A., Muhammad, K., Ding, W., Palade, V., Haq, I. U., & Baik, S. W. (2021). Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. Applied Soft Computing, 103, 107102.

[11] R. Zhao, H. Ali, P. Van der Smagt, Two-stream RNN/CNN for action recognition in 3D videos, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 4260–4267.

[12] M. Majd, R. Safabakhsh, Correlational convolutional LSTM for human action recognition, Neurocomputing 396 (2020) 224–229.

[13] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE access, 6, 1155-1166.

[14] A. Dosovitskiy, et al., Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.

[15] Noor, N., & Park, I. K. (2023). A lightweight skeleton-based 3D-CNN for real-time fall detection and action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2179-2188).

[16] Lin, L., Wu, J., An, R., Ma, S., Zhao, K., & Ding, H. (2024). LIMUNet: A Lightweight Neural Network for Human Activity Recognition Using Smartwatches. Applied Sciences, 14(22), 10515.

[17] Gupta, K., Singh, A., Yeduri, S. R., Srinivas, M. B., & Cenkeramaddi, L. R. (2023). Hand gestures recognition using edge computing system based on vision transformer and lightweight CNN. Journal of Ambient Intelligence and Humanized Computing, 14(3), 2601-2615.

[18] Huan, S., Wang, Z., Wang, X., Wu, L., Yang, X., Huang, H., & Dai, G. E. (2023). A lightweight hybrid vision transformer network for radar-based human activity recognition. Scientific Reports, 13(1), 17996.

[19] Choudhury, N. A., & Soni, B. (2023). An efficient and lightweight deep learning model for human activity recognition on raw sensor data in uncontrolled environment. IEEE Sensors Journal, 23(20), 25579-25586.

[20] Lizo, L., & Estrada, J. (2024, July). Lightweight Convolutional Neural Network (CNN) and Long Short-Term Memory Network (LSTM) for Dynamic Hand Gesture Recognition. In 2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA) (pp. 113-118). IEEE.

[21] Verma, K. K., Singh, B. M., & Dixit, A. (2022). A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. International Journal of Information Technology, 14(1), 397-410.

[22] Vrskova, R., Hudec, R., Kamencay, P., & Sykora, P. (2022). Human activity classification using the 3DCNN architecture. Applied Sciences, 12(2), 931.

[23] Verma, K. K., Singh, B. M., Mandoria, H. L., & Chauhan, P. (2020). Two-stage human activity recognition using 2D-ConvNet. IJIMAI, 6(2), 125-135.

[24] AlMuhaideb, S., AlAbdulkarim, L., AlShahrani, D. M., AlDhubaib, H., & AlSadoun, D. E. (2024). Achieving more with less: A lightweight deep learning solution for advanced human activity recognition (har). Sensors, 24(16), 5436.

[25] Basly, H., Ouarda, W., Sayadi, F. E., Ouni, B., & Alimi, A. M. (2022). DTR-HAR: deep temporal residual representation for human activity recognition. The Visual Computer, 38(3), 993-1013.
[26] Verma, K. K., & Singh, B. M. (2021, November). Vision based human activity recognition using deep transfer learning and support vector machine. In 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) (pp. 1-9). IEEE.
[27] Verma, P., Sah, A., & Srivastava, R. (2020). Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition. Multimedia Systems, 26(6), 671-685.
[28] Shafizadegan, F., Naghsh-Nilchi, A. R., & Shabaninia, E. (2024). Multimodal vision-based human action recognition using deep learning: A review. Artificial Intelligence Review, 57(7), 178.
[29] Rehman, S. U., Yasin, A. U., Ul Haq, E., Ali, M., Kim, J., & Mehmood, A. (2024). Enhancing Human Activity Recognition through Integrated Multimodal Analysis: A Focus on RGB Imaging, Skeletal Tracking, and Pose Estimation. Sensors, 24(14), 4646.
[30] Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.
[31] K.-I. Funahashi and Y. Nakamura, "Approximation of dynamical systems by continuous time recurrent neural networks, Neural Netw"., vol. 6, no. 6, pp. 801-806, 1993.
[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.
[33] S. Siami-Namini, N. Tavakoli, A.S. Namin, The performance of LSTM and BiLSTM in forecasting time series, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 3285–3292.
[34] Murad, A., & Pyun, J. Y. (2017). Deep recurrent neural networks for human activity recognition. Sensors, 17(11), 2556.
[35] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555
[36] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6), 602–610.
[37] Neverova, N., Wolf, C., Taylor, G., & Nebout, F. (2015). ModDrop: Adaptive multi-modal gesture recognition. IEEE TPAMI, 38(8), 1692–1706.
[38] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term Recurrent Convolutional Networks for Visual Recognition and Description. CVPR.
[39] Dosovitskiy, A. et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929
[40] Xia, L., Chen, C. C., & Aggarwal, J. K. (2012, June). View invariant human action recognition using histograms of 3d joints. In 2012 IEEE computer society conference on computer vision and pattern recognition workshops (pp. 20-27). IEEE.
[41] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).
[42] Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 6546-6555).
[43] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).
[44] C. Wang, Y. Wang, and A. L. Yuille, "Mining 3-D key-pose-motifs for action recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 2639–2647.
[45] R. Anirudh, P. Turaga, J. Su, and A. Srivastava,"Elastic functional coding of human actions: From vector-fields to latent variables," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3147–3155.
[46] J. Liu, A. Shahroudy, D. Xu, A. K. Chichung, and G. Wang, "Skeletonbased action recognition using spatio-temporal LSTM network with trust gates," IEEE Trans. Pattern Anal. Mach. Intell., 2017.
[47] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," in Proc. IEEE Winter Conf. Appl. Comput. Vis., Mar. 2017, pp. 148–157.
[48] Z. Liu, C. Zhang, Y. Tian, "3D-based deep convolutional neural network for action recognition with depth sequences," in Image Vis. Comput., vol. 55, pp. 93–100, 2015.
[49] Verma, K. K., & Singh, B. M. (2021). Deep multi-model fusion for human activity recognition using evolutionary algorithms.

Authors' Profiles

Vijay Singh Rana is a research scholar in Department of Computer Applications, College of Smart Computing, COER University Roorkee. He is an active researcher with experience in the field of Computer Science and Engineering. He holds an MCA and M.Tech(Computer Engineering) and is currently serving as an Assistant Professor at J.V. Jain College Saharanpur India. His research area includes Artificial Intelligence (AI), Machine Learning(ML), Gesture and Activity Recognition etc.

Ankush Joshi is an accomplished academic and researcher with over 14 years of experience in the field of Computer Science and Engineering. He holds an MCA, MTech (CSE), and a Ph.D. (CSE), and is currently serving as an Assistant Professor at COER University, Roorkee, India. Dr. Joshi specializes in Artificial Intelligence (AI), Machine Learning (ML), and Data Analysis, with a strong focus on interdisciplinary applications of these technologies. He has made significant contributions to academia, with more than 20 research papers and book chapters to his credit. Dr. Joshi has edited one book with IGI Global Publications and authored another, showcasing his expertise in the field. He is an active member of the academic community, having served as a reviewer for several IEEE conferences

and IGI Global publications. Additionally, he has chaired sessions at prestigious IEEE and Springer conferences, further solidifying his reputation as a thought leader in the domain of computer science and engineering. His work continues to inspire students and researchers alike, bridging the gap between theoretical knowledge and practical applications in AI, ML, and data analysis.

Kamal Kant Verma is working as Professor in School of Computer Science and Engineering at IILM University Greater Noida India. He did PhD in Computer Science and Engineering from Uttarakhand Technical University Dehradun India in 2021. He did B. Tech in Information Technology in 2006, MTech in CSE in 2012. He has 18+ years of teaching and research experience. His research area is Human Activity Recognition, Human Computer Interface, Pattern Recognition and Signal Processing. He has published more than 40 research papers in reputed national/ international journal and conferences such as Springer, Elsevier, IJIMAI, etc.

How to cite this paper: Vijay Singh Rana, Ankush Joshi, Kamal Kant Verma, "Lightweight 3DCNN-BiLSTM Model for Human Activity Recognition using Fusion of RGBD Video Sequences", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.6, pp.176-193, 2025. DOI:10.5815/ijitcs.2025.06.10